AUSTRIAN JOURNAL OF STATISTICS
Volume 28 (1999), Number 2, 49–79

Metadata Management in Official Statistics–An IT-based Methodology Approach

Karl A. Froeschl
University of Vienna, AUSTRIA

Abstract: There is a considerable impact of evolving information technology (IT) on Official Statistics. Coping with both risks and potentials of this development, Official Statistics is in serious need to address questions of keeping the statistical system functional and responsive to increasing requirements whilst facing rather tight economic conditions. At any rate, it can be anticipated that new approaches to information management will play a prominent role in the re-engineering of statistical infrastructures; in particular, metadata management will turn out as *key factor* in this indispensable renewal of statistical data production and dissemination systems. After tracing emergent developments back to their interwoven origins in both statistics and IT hemispheres, this paper singles out the dominant issues and components of next-generation metadata-based IT frameworks taking shape in the domain of Official Statistics, and proposes some framework design principles.

Zusammenfassung: Die amtliche Statistik unterliegt als qualifizierter *Informationsdienstleister* naturgemäß erheblich dem sich verändernden Produktionsmittel „Informationstechnologien“ mit all seinen Auswirkungen auf das Produkt „statistische Information“ selbst. Mehr noch als viele andere Bereiche der öffentlichen Verwaltung betrifft die amtliche Statistik die Frage des Umgangs mit den Risiken und Potentialen der IT, um einerseits die Funktionalität des statistischen Systems zu sichern bzw. den wachsenden Anforderungen entsprechend auszubauen und andererseits auf die zunehmend restriktiven Etatbedingungen strategisch zu reagieren. Innerhalb der sich wandelnden Bedürfnisse auf der einen und den technologischen und organisatorischen Möglichkeiten auf der anderen Seite gilt es, Szenarios für die amtliche Statistik der nächsten Generation zu entwerfen. In jedem Fall werden neue Ansätze des Datenmanagements im Brennpunkt der informationellen Reorganisation statistischer Systeme stehen, vor allem stellt sich das *Metadaten-Management* - die Strukturierung statistischer Datenbestände in Verbindung mit formalisiertem inhaltlichem und produktionsbezogenem Wissen - als zentrale Aufgabenstellung jedes effizienten statistischen Informationsmanagements heraus.

Keywords: Information Management, Meta-information, Statistical Information Systems, Official Statistics.

1 Official Statistics and Information Management

In reflecting major issues in regard of the organization of up-to-date statistical information systems and their design principles, this treatise covers three interrelated themes, viz.

- a review of more or less obvious impacts of current and anticipated information technology (IT) developments on institutional organizations in the domain of Official Statistics;

- the exploration of an IT scenario for Official Statistics helping to maintain or redefine the functional roles of established players in the field in a rapidly evolving environment;

- the rather exciting opportunities of adaptive institutions coping effectively with the challenges ahead, as opposed to the indeed gloomy expectations for statistical offices not recognizing in time the developmental dynamics of ongoing transitions.

It is a basic premise of what follows that evolving IT is both, driving the changes forcing Official Statistics to take responsive action, and providing the means to actually respond. To do so adequately, however, some *strategic perspective* as well as a set of *sound principles* is needed to turn recognized needs into well-targeted measures. Not denying other aspects intruding upon Official Statistics, this presentation will focus particularly on the role advanced IT can play in supplying an encompassing technical business framework for the conduct of institutional statistics. In all that, the discussion revolves around the pivotal notion of *metadata* (cf. Grossmann, 1999, for instance). Originating from various contexts such as statistical documentation in on-line systems and, later, statistical expert systems (Hand, 1994), this term meanwhile has become a crystallizing focus in statistical information systems development. It is likely going to be used synonymously to the latter as, apparently, next generation statistical information systems will inevitably have metadata at their very heart. In fact, the principles of metadata management are about to evoke a thorough reshaping of the statistics information landscape, expectedly keeping little of traditional organization structures and procedures. Thus, the real point is not how to make use of new technologies in Official Statistics–rather, it is what new technologies will make of Official Statistics before long.

In this section of the paper, after reviewing major determinants currently driving changes in Official Statistics and summarizing presently pursued measures, a lack of theoretical underpinnings for devising *statistical IT frameworks* is stated. The remaining Subsections 1.2 and 1.3 of this section then scrutinize, in some detail, the general impact of IT development on the statistics business as well as the changing role of Official Statistics in the emanating Information Society, in either case re-affirming *meta-information management*–the management of information necessary to produce and make use of statistical information–as the pivotal notion to which technical options and business demands converge. Recognizing this key position of meta-information, Section 2 highlights the main pillars of any formalized IT-based meta-information management: modeling the information processing chain and statistical data modeling. Section 2 concludes arguing that–on condition of a sufficiently formal approach to meta-information modeling–processes and data representations can be interlocked effectively such that information processing, and data production in particular, can be controlled to a good deal by *formal* metadata–thus transforming formalized meta-information into a first-rate data production and management resource. On top of this, Section 3 of the paper makes some proposals on the design of integrated metadata management solutions, turning formerly isolated

information systems and data holdings into subsystems of a unified "wide-area" statistical information infrastructure, and specifically addresses issues of systems interfacing for the sake of data combination. The concluding section links recent domain issues to the proposed methodology and tries to assess briefly its longer-term implications to statistical information management. Clearly, as it stands this methodology proposal (like any other around, by the way) is far from the maturity required to guide real metadata management system development. Rather, these considerations contribute only bits and pieces to a research program aiming at a coherent methodological framework which, of course, cannot be accomplished without a close linkage between theory and practice.

1.1 Official Statistics in Change

That reality, and especially socio-economic reality, undergoes rapid change these days is a commonplace observation. New political and societal contingencies (EU, globalizing economies, the Information Society, ...), innovations in information and communication technologies (telematics, inter-operable systems, information highway, multimedia, ...), and all of these accompanied by a host of new challenges and opportunities, dynamically permeate virtually every aspect of life. Against this background, the operation of Official Statistics–the main "reporter" about economic and societal change–is affected tremendously in two ways, viz.

- on the one hand, society as a whole and political administrations in particular put ever higher demands on statistical services (without willing or being capable to boost resources proportionally) while,

- on the other hand, both the provision of statistical services itself and the efficiency of their generation depend crucially on exactly this IT development pushing overall change.

This is to say that Official Statistics is forced to respond, at the same time, to new requirements regarding product content and quality (timeliness, accuracy, level of detail in supplied quantitative information, tender-based deliveries, etc.), to improve economic task fulfillment, to address new "European" demands (in case of EU-based institutions), and to straighten up internal production resources and methodologies. Furthermore, statistical offices increasingly find themselves in a reactive, market-driven position rather than operating any longer as a governmental, economically more or less sheltered authority.

Therefore, though with varying impetus, a re-engineering of statistical data production and information processing is generally taking place in Official Statistics (e.g., cf. CES, 1999). Most players in the field are now keenly adopting new information technologies (such as client/server architectures, workflow systems, on-line databases, Internet, Intranet, etc.), mainly for the sake of improving internal production efficiency and raising customer satisfaction by substituting off-line with on-line operations and services. The advent of the *Internet* can be cited probably as the most outstanding single factor affecting the present appearance of and the expectations set in Official Statistics. In this respect, many offices have begun to reconsider their Internet information dissemination policies, and this endeavor quite naturally has further accentuated the topic of metadata

and, particularly, the need for effective metadata management approaches. Essentially, the measures taken revolve around grading up established data dissemination services with on-line documentation, thus basically transferring accustomed documentation standards from (typically) printed publications over to electronic media. Much activity in this direction can be observed at individual statistical offices and (international) institutions with a statistical involvement alike.

However, these mainly reactive measures make more fundamental difficulties come to the fore. In order to fully utilize the potential of digital media and meet rising expectations, a decidedly more profound and proactive approach to reorganization becomes inevitable–it turns out that no less than the whole statistical information management needs to be reconsidered. This might be highlighted by just two elementary observations:

- After more than a century of automation in counting and numerical computing (thus lowering marginal operation cost close to none), awareness shifts to real expenses incurred from data management; consequently, data management economics–replacing informal modes of information management with an IT-based *rational methodology*–gains prominence.

- After nearly two decades of personal computing with a tremendous decrease of hardware prices and the general availability of powerful software, computing is no privilege of data producers anymore–everyone wishing to carry out statistical analyses of whatever sophistication can do so virtually unlimited, provided that well-documented, good quality data is accessible conveniently.

Thus, rather fresh inroads to the entire business are required to manage this severe structural change: Official Statistics must utilize information technologies *strategically* for its own organizational development. The pivotal issue of this strategy consists in realizing that a sustainable generation of high-quality information services cannot be brought about by eventually adding extensive documentation to output data. Rather can data quality and documentation be achieved reasonably and economically through a data production organization genuinely *controlled* by metadata (cf. Bethlehem et al., 1999). As a natural consequence, this leads to comprehensive metadata-based statistical information management frameworks presupposing, in turn, some purposive meta-information modeling. Apparently, so long as the role and importance of a formal approach to meta-information modeling is not recognized properly, the vast potential of metadata as a data production *tool* of its own cannot be exploited practically. Hence, on the road towards such a framework the *formal modeling of statistical information management*–its mapping to metadata structures, actually–represents the decisive methodology.

To date, no such systematic and encompassing theoretical framework for statistical information management is available. Traditionally, investigations and proposals have reflected endeavors and problems of individual statistical offices/agencies rather than aiming at general models or theories. Could be, though, that theoretical efforts haven't had good (enough) reputation or seemed to fall short of contributing tangibly to practical solutions. While the necessity of theoretical underpinnings is now receiving wider acknowledgment and, hence, their lack is felt more pressingly (cf. Krug, 1998, as an example), theory development still lags behind–partly because of the scarce attention (compared to,

say, commercially/enterprise oriented information systems development) of the domain in academia. Despite noteworthy efforts (such as the research promoted by EUROSTAT), practical and theoretical developments in statistical information management have not yet cross-fertilized satisfactorily. Accordingly, there is a lot of reason for yet another proposal like the one discussed in the following paragraphs of this paper.

1.2 The Impact of New Media and Technologies

Official Statistics and "new" information technologies in fact got married very early. Driven mostly by uprising difficulties in the dealing with census evaluations, it was the U.S. Bureau of the Census seeking innovative approaches towards the end of the $19th$ century to what later became known as *data processing*. This engagement led to the adoption of contemporary industrial "mass production" principles, notably HERMAN HOLLERITH's punched-card controlled devices, speeding up the Bureau's tabulations from the 1880 Census onwards. These successful automation ideas spread quickly and, indisputably, have left a lasting influence on the practical organization of institutional statistics to be felt still.

In technological terms, nowadays the situation seems reversed rather–it is the informatics development pushing Official Statistics. While in the days before the advent of commercial computing governmental policies had set the stage (this happened again when the U.S. Bureau of the Census encouraged the move to electronic computing by end of World War II), most of today's statistical institutions struggle with the pace of technological innovation. The apparent reason for this is the changed nature of IT itself: had former computing technologies (as rather passive tools like punched-card and later electronic computing) affected mostly the "data shop floor" of the business, IT nowadays actively concerns the *business models* of Official Statistics–as its core resource without which in fact entire statistical organizations are no longer thinkable. The view of venerable office departments using machines and software to carry out specific tasks gives way swiftly to the image of an information management framework constituting the backbone of a statistical organization and determining, in turn, what its departments–or, rather, operational units–are and how they cooperate best. This framework includes internal *processing flows* as well. Clearly, interactivity, personal computing, networking, and telematics applications involve more than just upgrading an office's computing center; rather, offices have to *reinvent* themselves bottom-up by technology integration. It is important to note, hence, that it is not the hardware or the programs that matter; quite on the contrary, the real impact of new technologies emerges from their socio-technical implications or, if you like, the *virtual organization structure* implied. As operational institution behavior and output become more and more a function of software (defining, in a sense, the "operation space" of an organization), material *meta-information management* becomes the prevalent paradigm of Official Statistics.

Accepting that Official Statistics becomes well-embedded in a ubiquitous information-technological landscape, its internal set-up needs to be fitted into it appropriately. This regards both the production structures and procedures as well as the interfaces to the surrounding informational environment. Changing requirements call, on the output side, for product designs specifically targeted at (post-) processing by external information con-

sumers of downloadable data/metadata. Quite contrary to traditional formats of information supply comprising a limited range of final products like published tables, press releases, self-contained printed statistical reports, and predefined on-line access modes for a small selected set of (mostly institutional) customers, now a wide and still widening range of dissemination channels and formats must be serviced (Zettl, 1997). Running different distribution media in parallel consistently places a heavy burden on information providers though, calling for an integrated approach to output organization embracing both data and documentation (metadata). Typically, multi-modal output management is conceived in terms of *data warehouses* embracing the compound production of a statistical institution, thus functionally separating data production from data consumption (rather than internal and external data usage).

However, viewing a statistical warehouse merely as a (central) output repository of a data producer must be considered a misconception. Rather, in order to reap the full warehouse potential it has to be placed right at the core of statistical information management in that *all* internal production processes are anchored in this pivotal resource. This way it becomes possible to get hold of all procedural and substantive relations between different production streams (such as surveys) run by the institution. The warehouse then provides a single homogeneous environment not only for all output-bound information processing but underlies all internal data management and processing routines as well. Additionally, as all in-bound processing feeds results directly into the warehouse, promoting IT-based solutions for data capture and data pre-processing suggests itself either, thus interfacing the infrastructure directly with respondent systems and administrative data sources (as proposed, for instance, by Keller and Ypma, 1997) as far as possible. Hence, suitably augmented, the warehouse concept in fact provides a functionally comprehensive carrier structure for all statistical in-house processing including, of course, any formalized meta-information management.

In regard of its linkage to the informational surrounding, certainly another important factor intruding upon the operational organization of Official Statistics is the increasing interlocking of offices and institutions with what might be called a network of networks. In fact, statistical offices are supposed to become major nodes in a global (statistical) information network collecting nearly as much information from other sites in the net as delivering back to it, suggesting the image of statistical agencies as *information-transforming intermediaries* between data production (sub-) networks and data consumption (sub-) networks. This becomes particularly apparent in the European context where (mostly) national statistical offices assume the role of information conveyors to EUROSTAT in addition to their normal operations. This “federated” (because of the sustained autonomy of the participants) model of distributed statistical information production sketches an upcoming unified architecture for statistical information management scaling, so-to-say, from micro- (within organizations) to macroscopic (between organizations) contexts, with apparent implications for inter-system communication and particularly metadata requirements. Preparing for an unbarred flow of statistical data between networked information systems implies an inter-operable system organization as well as highly standardized interfaces, both attainable and maintainable economically only on condition of a sufficiently elaborated framework of formally managed meta-information guiding metadata management in all participating local systems.

1.3 Value-added Information Services

Three broad functional domains in the generation of information services in the Official Statistics domain can be singled out, viz. the *production* of statistics services, their *utilization* (dissemination/consumption), and the generation of *added value*. Evidently, the "core business" of Official Statistics is data production. For historic reasons, data production used to be survey-centric with an input-oriented organization usually split into internal operational units according to either a substantive or a functional labor division principle. Typically, surveys are produced in a "pipelined" fashion passing on intermediary processing states from stage to stage along a predefined sequence up to final outputs, without any particular internal necessity to explicate accompanying meta-information such as regarding survey background or production details. In contrast to this traditional production view, the utilization/dissemination view reflects more explicitly external data consumer requirements in both product design and delivery policies, like improved and expanded production and background documentation, an enhanced range of products at various processing stages ("semi-finished" products, so-to-say), cross-survey data access and linkage (data harmonization and combination), and other customer-specific information services. Put simply, output orientation entails an output organization prepared to service a broad range of information inquiries definitely outside a producer's control.

Presently, there is an irrevocable shift from *input-* towards *output-orientation* underway in Official Statistics. Expectedly, the bureaucratic, low-adaptable model of "discrete" self-contained information production–prone to growing public suspicion because of its discreteness (Porter, 1996), by the way–will give way to business process models reflecting information-economic considerations in putting high emphasis on production structure and efficiency. Substantive output-orientation meets with technical requirements (cf. Subsection 1.2) here in that these tendency enforces the establishment of internal clearing-houses (Sundgren, 1997, Colledge, 1999) gathering and *integrating* the net gain of all subordinate "production lines". More or less, this kind of information integration amounts to present the compound output of a statistical institution in a *semantically mediated* way (that is, by metadata), laying open–to a certain degree–the multifarious substantive interrelations between different parts of the statistical information offer and making both offer and substantive relations amenable to formal exploration (like information browsing, navigation, enhanced retrieval modes such as metadata-based information access, etc.). Reasonable maintenance of such kind of clearing-houses implies an internal production organization allowing to directly feed-forward data documentation from the production process jointly with produced output data. This once more stresses the necessity of (re-)engineering production processes in terms of meta-information management principles so as to assure a coherent and consistent flow, or "throughput", of metadata alongside the processed data.

In a broader perspective, the tendency towards increasing output-orientation in Official Statistics will lead to a radically different view of statistical information management altogether. With exception of the initial production of raw data, the cardinal function of statistical data production consists in providing some information service generating–by way of input transformation–added "utility". If, now, the view of self-contained information production (as typical for the survey-centric style of production organization) is

abandoned in favor of this *service provision* view, each service can be framed up in terms of a transformation step converting input to output information. Accordingly, attention shifts to the genuine constituents of the production chain, sorting out functional units irrespective of whether they used to be "packaged" to single institutions or distributed among a couple of linked autonomous organizations. In line with such a generalized view, statistical processing is resolved into a stepwise transformation process of inputs, adding information utility step by step. In information-economic terms (Parker and Benson, 1988), hence, statistical service provision amounts to generate *added value* (although not, in general, representing a monetary or financial value). Since also the input-oriented single-survey processing view is amenable to such a functional decomposition eventually, a uniform theoretical concept of statistical data processing emerges deconstructing any statistical processing into transformation chains (directed acyclic graphs, technically speaking) composed of elementary value-adding transformation steps.

Resorting to such a purely *functional* statistics information production model (as opposed to the institutional one), now the vertical integration of service generation stages–or, depending on the point of view, distribution of transforming functions–becomes more a matter of natural responsibilities, economics, and interfaces rather than tradition and legislation (although, of course, data privacy regulations and disclosure control policies will exert a considerable impact still). Apparently, this view complies quite well with the previously stated 'network-of-networks' notion, and leaves ample room for various self-organizing schemes of cooperation between producers, consumers, and intermediaries. In particular, the added value business perspective suggests an aspect of statistical information management not yet recognized widely: by virtue of their outstanding position, statistical offices are ideally suited to assume the role of statistics *information brokering* as a specific value-adding service. While, for obvious reasons, the production of "base statistics" will always remain a core objective of Official Statistics, bringing together statistical information providers and information seekers in a (non-directed) information interchange network is about to become a tremendously important business. Moreover, in addition to the plain (yet in itself valuable) brokering role, "added-value" brokering could include the specific provision of *data harmonization and combination services*. A significantly boosted demand for such qualified mediating services will no doubt be brought in the wake of the upcoming Information Society. Because of their long-lasting involvement and profound subject competence established statistical institutions are in an excellent position to offer such services at both national and supranational levels.

The technical foundation of the sketched devolved statistical information management model is highly formalized information flows establishing a logical data production and dissemination infrastructure embracing all functional units involved. Advantageously, these information flows are based on information-controlled information processing presupposing, in turn, strict models of metadata management. In particular, information exchanged between functional units–whether within or between institutional actors–always comprises both data and metadata layers whence interfaces between functional units become, first of all, *metadata* interfaces. Therefore, the salient preparatory step enabling the envisaged value-adding statistical information processing is the development of a comprehensive and *institution-independent* metadata management framework. It can thus be justly claimed that metadata management issues are ranking among today's topmost chal-

lenges of Official Statistics.

2 From Data Buffering to Information Mediation

The administration of *object data* ("Sachdaten" in German)–that is, primary observation or register data obtained of respondents or administrative sources–can be considered the ancestral concern of data management in Official Statistics. Since long, the basic building block of this administration has been, of course, a database system. In an information systems perspective, these statistical databases store object data of all processing stages (raw data, preprocessed data, aggregates, final products), but doing so neither relating information units in subject-matter contexts–as needed for establishing clearing houses and integrated output databases–nor linking them alongside the processing flow–as required for a reasonable backbone structure to map workflow onto. Basically, this is a somewhat natural consequence of introducing databases as back-end *data buffers* (in between processing stages). Correspondingly, data processing has remained a "closed shop" business with quite restricted extramural database access, rather limited indexing facilities–typically, information units correspond to files identified by name or code–and rudimentary, at best, content documentation.

Data organization in Official Statistics used to be driven by technical design principles. For instance, locating an interesting piece of information amounts to navigate through hierarchical file directories rather than following *semantic* links to relevant substantive conceptual information units held in the data repository. Thus, information retrieval depends largely on human knowledge (about the accessed data holding), experience, and patience. From the production point of view, this kind of data organization implies that any information *about* processed data must be passed on informally and "manually" between successive processing stages. This makes labor division and function allocation schemes quite inflexible, costly, and–even worse perhaps–procedures tend to become fairly clumsy, slow, and opaque to extramural (and often even to intramural) data users. Moreover, data management necessitates human intervention in any but the most trivial data access operations, and dealing with specific information requests involves the ad hoc creation and execution of programs, a job normally only few specialists are available for. Apparently, such kind of data management is not amenable to any type of effective information interfacing across systems borders either.

These obvious shortcomings in data management have been analyzed quite intensely in the last couple of years, with North American and Scandinavian statistical offices taking a leading role (for instance, cf. Gillman and Appel, 1994, Graves et al., 1993, Sundgren, 1973, 1997). While, to date, this analysis has stimulated various tentative considerations (with little practical impact though), its main implication suggests a decided move towards *integrated statistical information infrastructures* based on a formalized dealing with statistical meta-information as outlined above. From a more theoretical, IT-inclined perspective, three pivotal aspects of–and, hence, major research areas in–modeling metadata-based statistical processing infrastructures can be singled out, viz.

- the development of a formal *workflow model for statistical information processing* (in particular, data production) in terms of functional transformations,

- the development of a *structural model of integrated statistical data/metadata holdings* as a state-space carrier underlying the workflow model, and

- the analysis of the *interplay* of both workflow and structural models having in mind the computational potential of information-controlled information processing.

Each of these topics is introduced briefly in the following subsections by stating the main themes, methodological approaches, and (preliminary) research findings, respectively.

2.1 Modeling the Processing Chain

Using the model of a service generation chain for statistics information production is a deliberate abstraction pursuing a double methodological purpose. On the one hand, this model supports (i) breaking down statistical processing into process stages separated by interfaces and (ii) delineating individual production components (see below). On the other hand, it helps to better grasp the requirements for metadata management and metadata interfacing in (re-)constructing the value generating transformation chains in terms of functionally separated process stages by sharply accentuating both metadata preconditions and returns of each of these stages.

To date, remarkably little material is accessible in the literature as to such a functional segmentation of statistics service generation, papers like van den Berg and de Feber (1992)–originating in the EUROSTAT-sponsored "Modeling Metadata" research project (Darius et al., 1993)–being certainly the most notable exception. While focusing on the processing flow of statistical surveys taken one at a time, that research highlighted especially the role of formal interfaces between successive processing stages: computationally, these become de-coupled by introducing intermediary "information objects" capturing as much meta-information passed on across stage transitions as possible. In other words, information objects could be viewed as "frozen states" of processing with a well-defined, agreed-upon metadata structure. In a functional abstraction, these interfaces mark *context borders* (implying that, within stages, mostly implicit contextual information is shared quite easily). In order to communicate the very meaning of object data beyond context borders, data context needs to be conveyed either. This is equivalent to communicate metadata and, thus, metadata becomes the prevalent means of context mediation across context borders (termed "context switching" by Bretherton, 1994, in a seminal draft paper on the topic).

Breaking down the workflow of statistical service generation (reflecting of course also the statistical data life cycle), three top-level processing phases turn up, viz.

(1) data production proper (survey design; data capture; raw data preprocessing; ingest of object data into some data register of cleaned, edited and often imputed object data ready for grossing and analytical data reduction),

(2) data mediation (distribution of data; "advanced" data production like tabulating, aggregating, grossing, statistical modeling, etc.; brokering, management of output information holdings, and provision of value-added services like data harmonization and data combination), and

(3) data consumption proper (terminal conversion of data into contingent empirical statements, usually including another layer of analytical data transformation).

In spite of their overlapping objectives each of these main phases entails its very specific processing context, stressing context borders–marked by data holdings in this case–and particularly the role of context switching. In principle, going down further the processing flow level by level, the same divide-and-conquer approach could be tried within each stage singled out. Yet, as processing steps at lower levels tend to *share* local contexts the interfacing metaphor is rendered decreasingly attractive. Rather, processing stages and steps contribute partial transformations centering round a shared set of object data and metadata, and only the joint application of several well-tuned steps yields the compound effect of a processing stage. Since choice, scheduling, and parameter initialization of the individual transformations making up a processing stage depend on its shared context, formal state representations in terms of metadata structures become predominant. In particular, combining object data of different sources calls for an integration of–as a rule, fairly heterogeneous–origin contexts expressed in terms of interface metadata: origin contexts have to be set in relation semantically such that meta-information gets merged into a shared context surrounding both data combination processing and the resulting combined object data. Technically speaking, context integration (materializing as *metadata combination*) is a prerequisite of data combination. Adopting this generalized view, shared contexts are local metadata management platforms into which input metadata is imported for object data undergoing transformation; on exit, transformed metadata is exported, or "projected", onto the output interface (standard) of the respective processing stage for subsequent context switching (Froeschl, 1999).

The lack of such mechanisms (and the corresponding interface standards) for metadata-managed context mediation is a major source of insufficiency, complaint, and frustration in present-day data usage at least in Official Statistics. Since the statistical data processing chain first and foremost is a metadata processing chain where input metadata is transformed into output metadata step-by-step (just like output data is obtained from input data), metadata must be thought of as inherently process-escorting information ("throughput"). Any conception of adding substantial amounts of metadata time and again to this process is entirely odd (this method is not only prone to inconsistency but hopelessly inefficient either). Moreover, it turns out that statistical meta-information divides functionally into *context switching* and *context sharing*, entailing metadata for bridging *changes in context* and for establishing *shared knowledge states*, respectively. While context switching metadata is message-based, and focusing on information transmission, context sharing metadata gives comprehensive descriptions of statistical data holdings and (transient) processing states.

A rather orthogonal view onto statistical service generation stresses the involved *production components* of statistical information. These designate the different transformation tools/steps which object data undergoes to render useful/comprehensible empirical information. Hence, among these production components one finds models of observational design, sampling, data capture, storage, data cleaning, editing, imputation, data linkage, bias compensation, estimation/grossing-up, disclosure control, aggregation, result presentation and layout, information dissemination, etc. In procedural terms, these components can be more or less identified with processing phases and stages. The

component-oriented analysis elicits critical meta-information details in that each production component presupposes specific metadata elements and structures from or contributes such to shared contexts. Furthermore, since components often require specific context elements originating at places far remote in the transformation sequence (for instance, grossing-up methods will be determined heavily by sample designs and data capture peculiarities, etc.), it helps defining context switching requirements as well.

While, in computational terms, this functional analysis of statistical information processing dynamics is focusing on a decomposition of workflow into elementary transformations, processing stages, state transitions, and state-space representations, the structural analysis–its counterpart–tries to reveal the semantic entities and relations required to formally *represent* meta-information ontologies.

2.2 Statistical Data and Metadata Modeling

Commonly, by 'data model' it is understood a formal symbol structure destined to capture salient aspects of a delineated set of real-world entities together with at least some of their relevant interrelations. The rationale of adopting a specific data model draws on its particular capability to express meaning by mere symbol arrangement; the "more" meaning a data model can purport this way, the better suited it is. A variety of approaches and techniques have been developed to devise data models and to put them into practice using established database management systems.

Since its inception, the *relational data model* and its derivatives–especially the so-called entity-relationship (ER and EER) modeling approach (e.g., Teorey, 1990)–have received particular attention because of their specific conceptual virtues: powered with formal uniformity, the relational data model combines mathematical rigor with clear-cut operational semantics by separating deliberately and crisply logical and physical representation layers. Over the last decades this has led to a wide adoption of relational databases also in statistical data management despite rather overt deficiencies of relational data modeling which provides just a basic carrier structure for the various "statistical relations" to be represented actually. Geared towards transaction processing based on tuple semantics, plain relational representation of data almost indispensably falls short of taking hold of the far more complex meaning of empirical data collections which are not just tuple assemblies (for one thing, observation records are tied together by a common sampling frame, etc.).

As no compelling alternative to relational modeling has appeared yet, there exist only preliminary and partial proposals for semantic data modeling in statistics. Primarily, so-called *macrodata*–structures purposively designed for aggregate representation–have caught attention, probably because of the more apparent mismatch of aggregate and tuple semantics. In this respect, Sundgren (1992) contributed early pioneering work-especially his '$\alpha\beta\gamma\tau$'-model–providing a high-level semantic analysis of statistical aggregate structures; the "Italian school"–centering around M. Rafanelli and F.L. Ricci–promotes an algebraic dealing with relational tables wrapped up in "statistical entities" treating whole datasets and aggregates (notably, statistical tables) as instances of abstract data types (Rafanelli and Ricci, 1993). Although these attempts, no doubt, move in the proper direction, they still remain confined to *relational* semantics and thus largely lack the ex-

pressive power required for self-contained statistical data modeling frameworks. Perhaps more promising looks a top-down modeling of business processes (as tackled by the U.S. Bureau of the Census; Gillman et al., 1996) because of its inherently "systemic" approach.

Statistical data modeling first and foremost is a task of statistical meta-information modeling. In order to properly reflect the very semantics of statistical datasets and their semantic relations among each other as well as to the (local and mediated) processing contexts, specially-devised *statistical data models* (cf. Froeschl, 1997, particularly Subsection 1.3) with considerable scope and expressiveness are called for. Among the meta-information these data models have to take care of there are

- the representation of *observation structures* (sample designs, variables, measuring units, scales, classification schemes, questionnaires, etc.),
- the representation of *aggregate structures* (multi-way tables, time series, indicators, etc.),
- the representation of *data production structures* (statistical units, sampling units, populations, masses, stratifications, weighting schemes, under-coverage, etc.), or
- the representation of *production component relationships* (linking production components both "transversally" to and alongside workflow),

to name just some. Of course, other approaches to partitioning meta-information–like, for example, into content-, method-, processing-, and administration-oriented meta-information (cf. Kent and Schuerhoff, 1997, with respect to a similar proposal)–are conceivable as well. In a rough distinction, oriented towards computational requirements, statistical data modeling comprises at least four–interrelated–areas of semantic relations to be mapped onto metadata structures, viz.

- *statistical data models in the narrower sense*, representing the internal structure and semantics of statistical datasets (at any level of processing ranging from "raw" case or source data to "final" output aggregates and analytical result figures like statistical indicators, etc.);
- operand structures encapsulating "integral" *statistical composites* consisting of some statistical object data set, accompanied by auxiliary numerical information (such as sample fractions, weighting parameters, etc.), and operand-specific metadata at various levels of formalization (from strictly formal to free text), all depending as to structure, degree of instantiation, and presence on the operand's respective processing stage (in technical terms, statistical entities are *typed* operands belonging to a class hierarchy of abstract data types);
- *shared contexts* (cf. Subsection 2.1) representing meta-information and meta-information relations essentially independent of existence and shape of statistical object data and composites in terms of (persistent) structured semantic networks establishing notions, entities, and terminologies of statistical subject-matter domains as well as the associations between domains, notions, entities, and terms (associations may be either implicit by name reference or explicit by formal linkage);

- *processing scripts* encoding (predefined, parameterized) sequences or subsequences of data transformations expressed in terms of elementary processing steps–"atomic transformators"–applied to (generic) statistical composites (much like computer programs are applied to their input data).

As additional metadata structures, temporary *analysis contexts* hold metadata elements (taken either from the underlying shared context or the processed statistical composites) needed for selecting, instantiating, or performing the transformations of a processing script. Although of auxiliary purpose only, local analysis contexts are indispensable for maintaining integrity and consistency of processed information by keeping, establishing, and updating the formal linkage of statistical composites to the shared context they belong to. Upon finishing the respective processing threads, analysis contexts expire after saving all metadata elements to be kept in shared context and/or statistical composites as necessary.

The specific virtue of this integrated meta-information modeling approach is its implicitly effected meta-information *standardization* (see also Subsection 4.1) which, in turn, is an indispensable prerequisite for meta-information interfacing as outlined in Subsection 2.1 above. Contrary to proposals emphasizing standardization of content-level metadata (for instance, Oleñski, 1996), however, statistical data modeling aims at a standardization of the *structure-level* of meta-information. In other words, it is not the particularities of individual subject domains (like a specific nomenclature, say) building the target of standardization but invariant, ever-recurrent metadata structures (like the generic format any nomenclature whatsoever shares by its nature). Thus, structural metadata standards can be identified with formal modeling languages providing a *framework* for expressing substantive metadata of whole subject domains (and beyond). These modeling frameworks encompass both, the (generic) entities to model and the (generic) relationships between entities. Practically, a framework may stack several generic model language/instance-pairs upon each other such as shown in Figure 1. This stacking implies that the instance of a higher-level modeling level is itself a modeling level of a yet lower-level instance. As an example, in setting up the description of a statistical observation frame, the general structure–at the top of the modeling hierarchy–is fixed by design while the specific observation frame is an instance of this general structure; yet, this specific observation frame determines (among other things, of course) a schema description for possible record formats of observation data, one or several of which, in turn, are the instances used for actual datasets.

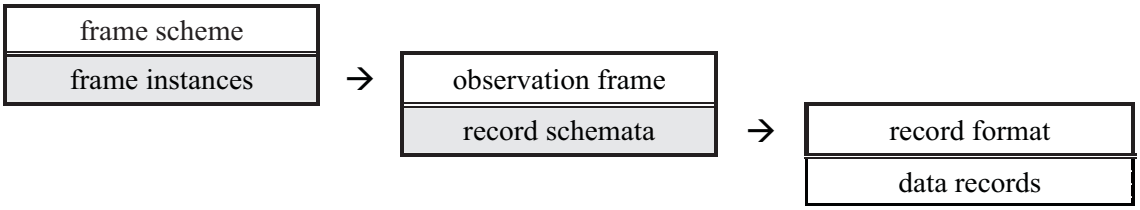

Figure 1: Example of Layered Metadata Structures

Much in the same spirit, for instance, also Sundgren's (1993) survey-centered data model–called "metaobject graph"–can be easily interpreted as a three-layered modeling stack. This generic survey model departs from the plain observation that typical institutional surveys are arranged as series of (mostly periodic) repetitions of some basic data recording scheme, named "survey occurrence". Thus, on the upper layer of the metaobject graph (stated as ER-diagram, by the way), starting from a fixed generic survey description model, a survey type instance is defined. This survey type instance provides the schema for a survey series instance, the different instances of which, in turn, represent the schemata for actual survey occurrences into which occurrence-specific information (data and metadata) is then placed. The hierarchical set-up in connection with increasing model refinements added top-down (metadata "granularity") suggests, of course, an incremental schema arrangement in the flavor of feature inheritance in object-oriented programming such that lower-level instances of the stack implicitly inherit structure and content elements of higher stack levels instances.

Regardless of any stacking, meta-information models are just data models–data structures encoding meta-information in terms of metadata; however, they are meta-data models in that they determine the structure of "instance" data models leaving considerable room for instance model variability. From a technical point of view, though, the standard schema-layer/instance-layer level-pair architecture (ISO, 1995) is perfectly sufficient for implementation since schema-layer data is processed like any other instance-layer data. Semantically, of course, the distinction is crucial because, by enabling unrestricted metadata processing this way, all object data-level operations taking place can be mirrored by meta-level operations, both allowing for an unprecedented flexibility in context representation and making provision for metadata-controlled information processing. This opens an interesting perspective for metadata management in arranging statistical information processing infrastructures as outlined in Section 3 below.

2.3 From Data Description to Embedded Metadata

Generally speaking, meta-information means data description, and metadata formally designates an excerpt of that meta-information. In a wide sense, meta-information encompasses *any* "piece of knowledge" of potential relevance for processing a dataset, or in interpreting it or the results derived from it. This, in fact, has stimulated a (still unfolding) plethora of metadata definitions (cf. Froeschl, 1997, for a brief overview), each with its own merits and shortcomings. However, without taking the meaning of meta-information beyond dispute, it clearly is neither useful nor tractable. A particular weakness common of most metadata approaches reported is the (usually tacit) assumption of the human addressee of metadata, that is: the purely documentary purpose of metadata. Although less obvious perhaps, metadata can do a much better service if structured appropriately. In particular, as already pointed out, rigorously formalized metadata is amenable to algorithmic processing just like ordinary data and, thus, can assume a driving role in process control; Bethlehem et al. (1999) have suggested to call this metadata "embedded".

The beneficial role of embedded metadata is twofold. Within shared contexts (see above), if statistical object data is wrapped up in statistical composites, attached metadata is open to processing jointly with object data. Hence, conceiving statistical composites

the operands proper undergoing transformations is an essential prerequisite to a continuous and consistent update of metadata elements. This concept, by the way, provides the computational basis for a generalized transformation calculus out of the building blocks of which statistical service generation chains can be synthesized as sketched in Subsection 1.3 above.

In case of context switching the role of embedded metadata is even more fundamental. If statistical composites cross context borders–as it happens in using some dataset not produced on one's own–it still carries its metadata (if it does so indeed), including various references back to its home context (such as a warehouse–cf. Subsection 1.2–it originates from). Contrary to the composite's proprietary metadata, however, there is no direct access to the referenced home context anymore. To this end, that home context must provide a metadata interface for its mediation to the outside. Traditionally, and reflecting current practice, such metadata interfaces (of contexts as well as individual datasets) are erected and maintained mostly manually, a laborious, resource-binding, and hardly-motivated activity leading, in general, to rather low quality and often blatantly inconsistent metadata. This comes as no surprise, though, given the difficulty and complexity of the task of documenting with hindsight the subject-matter background and the manipulations object data received all the way down, often under someone else's control. The decisive advantage of a formal metadata management in this respect becomes immediately apparent upon realizing that metadata interfaces are just *views* on metadata structures: views–a notion borrowed from the database domain–are metadata structures of their own but derived computationally from shared contexts. This is to say that view arrangement involves a functional specification rather than a substantive compilation of interface metadata, yielding tremendous gains in efficiency (replacing metadata production with metadata transformation) and flexibility (changing views amounts to adapt functional view specifications only). In addition to its economic leverage, this approach to metadata interfacing contributes significantly to metadata quality either since all views are derived from the same shared context metadata (that is, views are consistent) and interfaces are complete in that any context element required in a view can be easily included (except perhaps for some extra metadata inserted by hand). By symmetry, interface metadata are integrated into target contexts on import, again using functional metadata transformations (cf. Subsection 3.2 below). This way, imported (downloaded) statistical composites can be embedded appropriately in their new shared context by mechanically redirecting the home context references in their attached metadata utilizing interface metadata.

A further effect, on condition of a sufficient formalization of metadata management, context switching becomes neatly integrated into statistical service generation chains. In particular, by mediating context functionally across context borders, value-adding transformations and context switching might be interchanged (almost) arbitrarily since all object data processing takes place "in context". Conversely, without attaining a substantial level of formalization, metadata approaches to statistical information processing are likely doomed to fail by and large: it is the *metadata processing* harnessed safeguarding the supply of metadata in a desirable quality only. The more innovative statistical offices have already recognized the far-reaching implications of metadata processing as to its potential for increasing both internal productivity and consumer service levels (for instance, Statistics Netherlands recently reported on an anticipated productivity gain by a factor of 50

over 10-15 years; cf. Kent, 1998).

Just to be sure, it should not be expected that formal metadata will ever do away with other, "weaker" types of metadata; in particular, a distinction between metadata of *deep* and *shallow* semantics seems useful. Still of formal nature, shallow metadata–capturing "flat" associative relations between entities only–provides the fact basis for information browsing (tables of content, dictionaries, directories, indexing facilities, etc.) whereas metadata with deep semantics faithfully reflect internal structure and meaning of denoted entities. As a residual case, entirely unstructured metadata–better termed *metatexts* since of no use for algorithmic processing–might be added for the sake of completeness and, if need be, for complementary documentation. Quite obviously, since formal metadata can always be converted into legible documentation but not vice versa, metatexts should be resorted to in (exceptional) cases when there is no way to express meta-information in formal terms. Computationally, metatexts documenting statistical object data are, at best, passed on as tags attached to statistical composites.

To date, the concept of embedded metadata has not yet had much impact on actual software systems development. Most notably, the idea has been utilized in the BLAISE data capture system (Schuerhoff, 1993), and was explored to quite some depth in the IDARESA system (Grossmann et al., 1998). That project aimed to provide machine support in deriving harmonized statistical aggregates in front of a logical federation of autonomous metadata-coupled statistical data holdings contributing information to comparable subject-matter domains. IDARESA research evidenced both soundness and power of the embedded metadata concept, in addition to devising a formal data transformation model encompassing all stages of statistical output processing (Denk, 1999).

3 Meta-information Design

Neither theoretical design concepts–such as those outlined–nor piles of functional requirements contribute a system *design methodology* in itself. Thus, regardless of comprehensive reassessments of established practices, tenets, and approaches of the field, the actual development of integrated statistical information processing frameworks resembles an artful skill, resting on not exactly abundant prior experience, and moving on by trial-and-error mainly.

In spite of this, perhaps some general indications for framework design can be stated based on the analysis presented in the previous sections of this paper. First of all, it immediately follows that meta-information modeling must not be limited to a particular stage of data processing, or to a particular branch or type of statistics, to unfold its full virtue. Rather, any meta-information model ought to encompass the whole processing context of a statistical agency, office, or unit, regardless of the actual arrangement of its computational infrastructure. In addition to this, recognizing the increasing importance of systems integration and coupling at various processing levels, the view of each component system having its own internal proprietary data model ought to give way to a perspective of component systems sharing a *common*, hence highly standardized, meta-information design. Although hardly ever that ideal, systems are then fitted in between deliberately placed interfaces rather than the other way round. Still this enables a considerable scope

of function integration schemes giving rise to and surrounded by shared contexts. This scope ranges from tight, stable integration of singled-out functions and production components, corresponding to more conventional information systems (like typical statistical data holdings currently in operation), to loose, temporary function integration owing its flavor more to telematics applications linking together "scattered" information systems in a dynamic *web of statistics*, so-to-say.

The following paragraphs sketch a suite of general meta-information design principles consistent with specifications and requirements as discussed in Sections 1 and 2. Whilst reflecting, to a large degree, experiences gathered in the already mentioned IDARESA project, the validity of these principles is endorsed by conclusions reached in other research projects of similar scope and orientation.

In what follows, the term 'statistical meta-information system' (e.g., cf. EUROSTAT, 1993) is avoided consciously because of its ambiguous meaning. Pronouncing this as '*meta-information* system'–an information system including also meta-information–amounts to regard such a system a functional part of an encompassing statistical processing framework (accordingly, any integrated statistical processing framework in the sense introduced is by definition a '*meta-information* system'). Conversely, a reading as 'meta-*information system*' assigns the utterly different meaning of "systems holding information *about* information systems". Taken this way, the term is indeed quite useful in that a statistical processing framework typically comprises one or even more such meta-*information systems*.

3.1 Design Principles

In order to design an overall statistical IT framework for metadata management into which component systems and interfaces are then built top down, a small set of well-tuned complementary basic principles is in place. Arguably, the most salient aspects of meta-information modeling with respect to requirements and design criteria discussed in the previous paragraphs can be condensed to three general principles, viz. the *terminology principle*, the *principle of representation integrity*, and the *principle of conceptual information access*. Each of these principles is characterized briefly in this subsection.

The Terminology Principle

The primary operands of statistical information processing–datasets of object data–are interrelated in multifarious ways by shared subject-matter terminology and concepts. It is tempting to use this naturally available "data language" and arrange information access through substantive terms (cf. Appel, 1993) instead of a technical, location-based retrieval (by, say, file names or the like). To this end, however, statistical discourse (or subject) domains first need to be resolved into carefully defined categories of data language terms such that these terms, as parts of object data descriptions, mediate a semantically well-defined information access (in the sense of shallow metadata as defined in Subsection 2.3). Moreover, since formal associations between terms can be used to reflect structural relationships between and within the object data carrying these terms (deep metadata semantics), a wealth of modes of formal exploration of a body of statistical information

comes into reach. This is the concern of the terminology principle resting on a formal structure model composed of

- a *meta-nomenclature* subdividing the (open-ended) universe of subject-matter terms exhaustively into a limited predefined set of term categories such that each term occurring gets assigned uniquely to a term category (this term is then said to be an instance-term of that term category), and

- a *meta-graph* defining a semantic network of (usually binary) associations–"meta-relations"–between term categories such that instance-terms of the term categories involved enter the respective instance-relations.

A dedicated language for statistics indeed, the meta-nomenclature provides a vocabulary of elementary distinctions recurrent in all statistical subject domains as constituting components of observation structures, aggregate structures, production structures, production components, etc. On top of the term categories of the meta-nomenclature, the meta-graph superimposes a network of labeled associations, this way formally expressing well-discerned semantic relations between the "meta-terms" linked. Jointly, meta-nomenclature and meta-graph, fixing logical metadata structures in a general yet ultimate way, determine which meta-information can be stated in terms of formal metadata eventually.

Setting up and maintaining the metadata structures–notably, shared contexts and statistical composites–of a subject domain now amounts to populate the term categories and meta-relations of the meta-graph with the respective instance-terms and term-to-term links of this particular domain. The resulting instance-graph then enables both, (i) content-oriented information access using instance-terms, by term category, as multi-dimensional searchable index and (ii) non-linear exploration–navigation–through the term space by picking associations of the graph's transitive closure. Likewise, thanks to the deep semantics encoded in the term associations, instance-graphs provide the backbone of algorithmic metadata processing (which the remaining design principles capitalize on).

"R&D Basic Pop"		BASICPOPULATION
attribute	*type*	*value*
ScopeRef	link	[] DOMAIN: "R&D Expenditures"
PopUnit	link	[DOMAIN: "R&D Expenditures"] BASICUNIT: "R&D Booking Entry"
PopType	code	(v)irtual
PrintName	text	R&D performing booking entry population
Description	text	This (fictious) population encompasses all booking entries in connection with R&D expenditures
Lookup	text	Frascati Manual (1993)

Figure 2: Attributes of Term Category "BASICPOPULATION" (Excerpt)

In order to illustrate the meta-graph concept, Figure 3 shows an excerpt of an instance-graph taken from a European R&D statistics domain implementation (reproduced from

Grossmann et al., 1998). The shaded nodes in this exhibit, bearing the actual instance-terms, represent individual (meta-) information entities (in the sense of integral storage units, or "objects"); on the upper side of each node the respective term's category is indicated. This excerpt considers the linkage of various substantive units and populations, including the labeled links between instance-terms. For example, "Research Sites" is an instance-term of category "SAMPLINGPOPULATION" representing a collective of individuals, identified with an instance-term named "Research Performing Unit" belonging to term category "SAMPLINGUNIT". The different, though formally related instance-terms create a (sub-) network of units and populations used to describe the observation structures of R&D financing and expenditure data more or less.

For the sake of simplicity, Figure 3 omits much detail from instance-graphs making them really useful. Actually, there are many more links (not necessarily visible at the human user interface level) in such a graph required to minutely capture deep semantics. Furthermore, depending on term category, to each instance-term is attached a pack of attributes recording various features of the entity denoted by this term. Figure 2 sketches an–abbreviated–attribute set for BASICPOPULATION "R&D Basic Pop" occurring in Figure 3. Obviously, these attributes comprise both formal and informal description elements; however, only the formal ones are amenable to algorithmic modes of information access and metadata processing (though informal attributes may still allow simple text-based retrieval like string search, etc.).

It might be worthwhile mentioning that there is an interplay between term attributes and meta-graph structure: actually, by decomposing ("deconstructing") meta-terms to yet more and more elementary components, meta-term structure gets elevated to meta-graph level. With regard to the layered metadata architecture depicted in Figure 1, this is tantamount to moving fixed content structure to upper schema levels, with a corresponding gain in instantiation flexibility and, hence, metadata modeling power. Thus, meta-nomenclature design is governed eventually by economic considerations aiming at a tenable trade-off between rich semantics of term topology and "compactness" of term representation. For example, it certainly is advisable to assign within-dataset relationships to the attribute dimension while between-dataset relationships are better represented at meta-graph level. Practically, reasonable solutions will often require repeated cycles of refinement and evaluation.

The Principle of Representation Integrity

The principle of representation integrity shifts emphasis to the formal representation of data production and data transformation, that is, to the process layer. In its essence, this principle amounts to rule out any manipulation of object data circumventing attached metadata. On condition that all statistical object data is wrapped in a "shielded" container structure such as statistical composites, any data access and data transformations are allowed to take place only mediated by metadata. Thus, thinking in technical terms of data abstraction, all operations on data (whether object or metadata) are functions defined on statistical composites implying that any transformation of object data (if any) can never become separated from the corresponding transformation of embedded metadata (cf. Subsection 2.3). Evidently, this "tight packaging" of both object and metadata

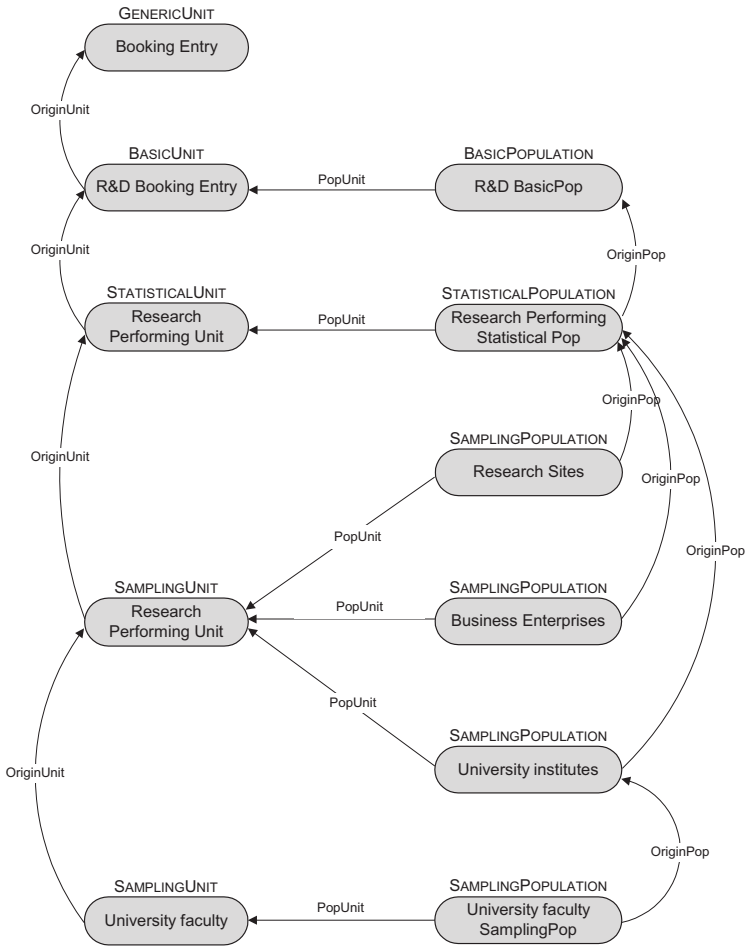

Figure 3: Excerpt of an Instant-graph

layers (together with further auxiliary information) safeguards both consistency and completeness of stored statistical composites. In a former research project–cf. Darius et al., 1993–these data/metadata packages, termed "tandems" then, have proven both manageable and powerful despite their initial limitations.

As a natural consequence, the principle of representation integrity entails a metadata throughput approach to statistical information processing from the very beginning of the data life cycle. In fact, since metadata already precedes actual object data in the early stages (such as survey planning or data capture preparation; cf. D'Angiolini et al., 1996), statistical data production starts with metadata processing rather than data processing–object data only comes in somewhere along the production course, getting tightly interlocked with its metadata from that point onwards. Thus, self-contained metadata processing chains necessarily originate from (and, by the way, always remain rooted in) data production contexts encoded formally in metadata structures.

Another aspect of the principle is the active role (Kent, 1998) of embedded metadata. From a technical point of view, this formal metadata can be regarded as a set of choices determining (part of the) meaning and sound usage of object data. Hence, these choices identify feasible data transformations while impeding others. This is of twofold relevance:

on the one hand, it enables a formal checking of technical applicability and semantic feasibility of intended data transformations; on the other hand, it alleviates the formulation of processing scripts (cf. Subsection 2.2) in that minor details of these can be left open to instantiation with metadata at invocation time. Considering that many metadata elements used in a transformation step are often generated by another step, this gracefully lets metadata management become facilitated mostly automatic by carrying out high-level processing scripts.

In a fashion similar to operand transformations, the principle of representation integrity also applies to shared contexts. Although less dynamically in general, shared contexts undergo transformation–state updates–either. In particular, in case of data combinations new shared contexts need to be established by merging together pre-existent origin contexts (cf. Subsection 2.1). If applied consistently, the principle enforces a formal, functionally defined context creation by, whenever possible, deriving its embedded metadata from the respective origin context counterparts. Likewise, endogenous state transitions of contexts (with minor exceptions regarding a manual insertion of metatexts perhaps) have to be triggered and effected by explicitly invoked metadata functions (mappings of metadata structures onto metadata structures). This principle also neatly integrates formal view management generating metadata interfaces at context borders.

Encompassing both, object data and data context, in a formal approach to metadata processing, the whole statistical service generation chains is mapped to an integral metadata management framework. However, still missing to a large degree are theoretical foundations of metadata processing, such as metadata combination and metadata reduction functions paralleling the respective object data level operations.

The Principle of Conceptual Information Access

Typically, traditional information retrieval inquires database extensions. This is to say that query responses are restricted to what is actually stored. However, especially with statistical databases, very often not the stored object data itself is of interest but the information to be gained from suitably transformed data. In logical terms, this calls for an extension of information retrieval to include what might be deduced–relative to some given set of deduction rules–from stored data. To this end, though, a (formal) language powerful enough to express a wide range of information "interests" is in place.

Like any other language, such a statistical query language consists of both grammar and vocabulary. Evidently, the bulk of vocabulary is provided with the instance-terms of domain terminologies whereas meta-nomenclature and meta-graph designs determine, by and large, language grammar. Quite importantly, such a language (like ***i*QL**; cf. Denk, 1999) abstracts from concrete data holding content in that the universe of formally feasible (query) statements cannot be straightly identified with stored data anymore. Rather, the totality of statements formable determines what can be queried in principle (Sato, 1989), without any assurance that a particular query statement leads to an appropriate response at all. Since query statements only denote content and structure of desired retrieval targets, derivability of responses in fact depends on (i) the availability of suitable base object data and (ii) a sufficiently "complete" derivation mechanism capable to transform selected base object data into what is circumscribed conceptually (but not procedurally)

in query statements. Considering that query statements themselves consist of pure metadata, the principle of conceptual information access entails a problem solving approach using metadata processing in order to figure out whether a query is answerable and, if so, which transformations must take place to generate a response actually.

In technical terms, processing "conceptual" queries for quantitative information resorts to a formal algebraic calculus (Froeschl, 1997) comprising a set of elementary operators for (meta-) data selection and transformation designed according to the representation integrity principle. Coarsely speaking, answer generation amounts to the task of synthesizing operation sequences (within the transitive closure of all sequences possible given the respective storage content of an inquired information holding) effectively transforming stored base data into the quantitative response denoted by a query statement. Because of the apparent combinatory nature of response generation, problem-solving heuristics–encoding well-proven derivation strategies and established data processing practices in predefined parameterized processing scripts–must be engaged to cut down search spaces and speed up synthesis of operation sequences. Still, though, query processing will mostly remain semi-automatic keeping in mind that a complete semantic formalization of statistical domains is out of question practically.

Whilst basically providing an output-oriented inferential user interface to statistical information holdings, metadata-based formal querying facilitates also the arrangement of enhanced system-to-system interfaces for context switching. Enabling a "smooth" interoperable systems coupling, conceptual information access is indeed a major prerequisite for establishing the envisaged statistical added-value generation view discussed in Subsection 1.3 above.

3.2 Interlacing Information

The functional approach to statistical service generation naturally emphasizes context interfaces (cf. Subsection 2.1). Generally speaking, two types of interfaces can be discerned, viz. vertical interfaces within processing chains stringing together successive processing stages, and horizontal interfaces between processing chains running "in parallel". Besides the apparent significance of vertical interfaces in processing chains, horizontal data combination interfaces receive growing interest (for example, in connection with the European Statistical System linking together the national statistical systems of the member states of the European Union). In fact, data combination has always been an integral part of (intramural) data production as even production of primary statistics involves the interlacing of auxiliary data such as, for instance, in sample drawing, for data imputation, or in adapting grossing weights to adjust estimates to certain distribution constraints, etc. However, in contrast to this production-bound data combination, horizontal data combination focuses on the consumption-bound interlacing of data originating from sources typically not intended or prepared for ever becoming combined. As statistical data repositories expand in scope and size, and data accessibility improves dramatically thanks to networked systems, it is indeed tempting to exploit the semantic overlap of empirical data regardless of its origin and squeeze out hidden information at little marginal cost. Similarly, data production, and especially data capture, is becoming more sensitive to both production effort and respondent burden, a tendency also promoting more considerate

data tapping schemes seeking to interlace existing sources (like administrative registers, etc.) to substitute for traditional survey data production (cf. Schulte Nordholt, 1998). All this suggests horizontal data combination as a topic of meta-information design and metadata modeling of its own.

In the database domain, recently quite many proposals have been made as to joining disparate database systems and database schemata (Sheth and Larson, 1990, Conrad, 1997). In particular, various approaches to integrating semantic data models underlying logical database designs have been explored, yet slowly bringing forth a well-founded "mediation theory" (Wiederhold and Genesereth, 1997) indispensable for the inter-operable coupling of information systems. Furthermore, because of their orientation towards transaction databases these considerations have little bearing on issues arising in statistical data integration and data combination.

Regarding the proposed integrated framework approach to statistical service generation, horizontal data combination amounts to merge distinct statistical composites originating from several home contexts (cf. Subsection 2.1). Acknowledging the metadata throughput premise of the framework, hence, metadata combination becomes the logical precursor to data combination. Apparently, in order to combine metadata formally, some operative platform (composed of metadata structures) is required onto which home contexts are mapped as far as needed for further processing. Depending on its persistency, such a platform hosts a shared context built either on demand from scratch (as part of a temporary analysis context) or as a stable metadata structure coupling all home contexts involved to what is incidentally called "tight federation" (the latter type was explored in the IDARESA project; Grossmann et al., 1998). Focal point of the combination platform is formal correspondence relations integrating attached home contexts as to both terminological and structural dimensions (Froeschl, 1999). By matching home contexts in a formal way, technical, terminological, and semantic overlaps and divergences are explicated as deep metadata fully amenable to metadata processing, thus providing a rigorous computational means for home context mediation.

Quite obviously, recognizing the immense know-how and effort needed, creation and maintenance of correspondence relations count among the most proficient added-value services of information intermediaries. This is of particular importance with respect to establishing harmonized statistics relevant in virtually every kind of comparative statistics or derivation of "global" aggregates fusing a variety of data sources. Traditionally, harmonization is aimed at by taking active measures in data context and production organization to eliminate home context heterogeneity from the outset ("pre-harmonization") imposing, in fact, a top-down implementation forcing each contributing data source involved to obey strict production rules and to stick to prescribed term definitions and concepts (such as nomenclatures, content standards, etc.). While quite promising in theory, pre-harmonization suffers from a variety of practical flaws and pragmatic disadvantages; for one thing, there may be very good reasons for keeping local context peculiarities, or changes over time, which simply will not vanish because of some centrally admonished harmonization stance. Accordingly, to be successful pre-harmonization must often give way to softer modes of compliance, leading to methods of "bottom-up" harmonization defining correspondence relations in retrospect, depending on purpose. In addition to leaving much autonomy to individual data sources, a framework design in deliberate

support of post-harmonization favorably admits setting up correspondence relations quite flexibly. Thus, assuming only modest levels of pre-harmonization, metadata-mediated post-harmonization is just fine for adaptively establishing data holding federations facilitating horizontal data combination, and data integration in general.

Adjoining a technical post-harmonization facility in a sense completes the design methodology of metadata-based statistical information processing infrastructures in that, virtually, any conceivable value-adding function integration is embraced. It also rounds off the design principles introduced in the previous subsection by expanding conceptual information access beyond the information offer comprised in a single data holding: referring to a combination platform, the "federated" offer of all its attached data holdings becomes accessible.

4 Linking to the Information Society, Conclusions

In developed societies Official Statistics is deemed a community resource neither exclusively nor specifically oriented towards the information needs of political representatives and administrations. At the European level (Franchet, 1995), statistics is regarded as vital for community integration (for example, economic convergence criteria, etc.). Particular attention is paid to comparative statistics of national accounts and socio-economic indicators, especially with respect to transition countries and nations applying for union membership. Another development of considerable impact on Official Statistics and particularly its current information dissemination policies are recent advances in information technology like the Internet. This broad range of tasks, aggravated by spurred social dynamics and globalization, tangibly challenges traditional organization structures and practices of Official Statistics. Remarkably, despite these tremendous changes, Official Statistics still quite often seems to reflect itself as a branch of political administration rather than advanced information business. However, Official Statistics will maintain and strengthen its role in the Information Society by offensive innovation only–aiming at IT-backed information management in charge of producing high-quality, demand-driven statistical information services for the whole community. In so doing, Official Statistics might even attain the role of a driving force in the field and, eventually, become itself a good deal of this often-cited information society.

In the concluding paragraphs, some topical issues and noteworthy initiatives towards improved utilization of information technology are addressed followed by a brief review of the broader implications of observable developments.

4.1 Information Quality, Standards, and Research

Among the topics recently receiving increasing attention is information quality, usually referred to as 'data quality'. Official Statistics gets more and more exposed to criticism as not to meet current and future quality demands satisfactorily. Obviously, information quality has many facets, and withstands a straight definition. While attached 'best quality' tags (regardless of the credibility of the issuing authority) will help little to improve it, information quality would become a rather self-evident property of statistical

data if data production and information processing as a whole complied transparently with reasonable process and structure standards. Hence, meta-information modeling is to be regarded instrumental for statistical information quality. A data receiver may decide about data quality directly by the embedded metadata incorporated in any processing of object data, without having to rely on other quality assessments or (even worse) on bulks of weakly-structured documentation difficult to verify and rationalize. Good quality data, then, basically depends on good quality metadata management.

To be useful at all, both data and metadata models must conform to standards. Standards, in turn, reflect requirements such as concerning communication vs. processing, or technical vs. structural vs. content-oriented issues. To date, a propensity of standardization efforts towards technical and communication-oriented standards is apparent. This ranges from electronic data interchange protocols and file formats (for example, cf. Boyko, 1999) to more informal "guidelines" prescribing wishful documentation elements attached to statistical data available on-line (METIS, 1998ab, OECD, 1996) or eliciting preferable modes of information delivery (Sutcliffe and Patel, 1998). Expectedly, activities in this direction will continue and intensify. In contrast to this, other areas of standardization are less vigorously pursued. A notable exception is the METIS long-term effort of the UN-ECE (Geneva) focusing on terminology and structure; in spite of providing a lively forum for the exchange of ideas, however, the METIS terminology (Prazenska, 1996) has failed to receive wider recognition most probably because it is too low level. More ambitious attempts towards structure standards deeply linked to internal business procedures are currently undertaken at the U.S. Bureau of the Census (Gillman et al., 1998). The least of all, though, process standards have been addressed. This is a serious omission since processes to a large degree determine the functional requirements of context and communication interfaces and, hence, data models as well (Froeschl, 1999).

The introduction of new technologies rarely takes place in one stroke. Rather, smooth and affordable ways of incorporating new technologies and systems gradually into grown operative frameworks must be looked for. If indeed the emergent re-engineering of Official Statistics information management goes ahead the way outlined, both intramural and extramural developments towards a unified meta-information management framework must proceed in a piecemeal fashion fitting IT infrastructures into the intended future shape of institutions.

One development already well underway in Official Statistics regards the refurbishing of traditional statistical data holdings, and their weaving into distributed networked statistical information environments. However, to date mostly technical developments have been given attention (such as Internet technologies, or the introduction of multi-tier client/server arrangements replacing mainframe-based IT infrastructures; cf. Sundgren, 1997) whereas the less obvious implications for meta-information management have received comparably low emphasis. This holds true for various research activities and programs including EUROSTAT's Development of Statistical Expert Systems (DOSES, 1989-1993) and Development of Statistical Information Systems (DOSIS, 1994-1998), and might be one of the reasons why real breakthroughs are long in coming. Accordingly, a severe necessity for yet more effort in metadata research is felt. It also remains to be seen whether current metadata research (again seeing EUROSTAT as a main promoter of technological development; cf. Ramprakash, 1998) and metadata standardization projects

will make significant IT contributions to Official Statistics information management. Success will depend on whether a widely supported consensus is reached as to the proper role of formal metadata within integrated statistical information processing environments, and whether adequate standards for metadata representation and transmission can be agreed upon. At any rate, it is fairly safe to bet that the topics of meta-information modeling and metadata management are here to stay for a while.

4.2 Some Broader Implications

Currently, Official Statistics is undergoing twofold reshaping. On the one hand, in a process of *inner* development the traditional institutional categories of data production and data consumption give way to a functional approach to statistical information services provision, while, on the other hand, a process of *outer* development responds to societal change and changing information demands. Evidently, both kinds of development intermingle tightly, and are separated here mainly for analytical purposes.

As to the internal developmental process, service provision is deconstructed into functional "modules" each of which both takes up and hands back information where 'information' materializes in statistical object data as well as metadata. In economic terms, these modules could be viewed as value adding components of a processing, or service production chain (cf. Subsection 1.3) whereas, technically speaking, they behave as information converters transforming input information into output information. The dissolution of service generation into modules facilitates a re-organization of statistical institutions by packaging modules to (possibly spatially distributed) functional clusters interconnected in a web-like communication structure spanning the whole life cycle of statistical information (cf. Subsection 1.2). All of this is mediated by digital technology fusing both communication and computation into a single uniform medium (cf. Subsection 1.1). Thus, functionally, differences of communication within institutions and between institutions tend to disappear, and information technology itself provides the statistical *transaction space* encompassing all involved actors (Conen and Neumann, 1997). Actually, in analogy to electronic trading and commerce, this transaction space is not only a repository of statistical information–it happens to be the basic platform for process organization and service generation proper. As technical prerequisites, meta-information modeling and the derivation of metadata structures and process standards (as outlined in Sections 2 and 3) become key development factors of the statistical information business.

Mainly, the inner development of Official Statistics is pulled by profound changes of the surrounding informational and socio-economic environment. Clearly, not to respond, or to react insufficiently, could deprive Official Statistics of its reputation and possibly even damage its position. However, fundamental change always bringing both threat and opportunity, Official Statistics–as its outer developmental strategy–might rather take a decidedly active part (as it did way back in HOLLERITH's days) in shaping the Information Society and move ahead using metadata management as a strategic device in renewing its mission.

References

G. Appel. Zum Entwurf eines metadatenbasierten statistischen Informationssystems. (In German). *Allgemeines Statistisches Archiv* 77: 68-91, 1993.

J. Bethlehem et al. On the Use of Metadata in Statistical Data Processing. Working Paper No. 23, UN-ECE/METIS, Work Session on Statistical Metadata, 11 pp., 1999.

E. Boyko. Statistical Meta-Data in Context: An Overview of Statistical Meta-Data and Related Meta-Data Systems. Working Paper No. 17, UN-ECE/METIS, Work Session on Statistical Metadata, 10 pp., 1999.

F. Bretherton. A Reference Model for Metadata–A Strawman. Draft Paper originally circulated as WWW document; cf. F.P. Bretherton and P.T. Singley. Metadata: A User's View. In J.C. French and H. Hinterberger, editors, *Scientific and Statistical Database Management* (Proc. 7*th* SSDBM), 166-174, 1994.

CES–Conference of European Statisticians. Report of the February 1999 Meeting on the Management of Statistical Information Technology. UN Economic and Social Council/Statistical Commission and Economic Commission for Europe, CES/AC.71/1999/31, 12 pp., 1999.

M. Colledge. Statistical Integration Through Metadata Management. *International Statistical Review* 67(1): 79-98, 1999.

W. Conen and G. Neumann, editors. Coordination Technology for Collaborative Applications–Organizations, Processes, and Agents. Springer (LNCS 1364), Berlin, 1997.

St. Conrad. Föderierte Datenbanksysteme: Konzepte der Datenintegration. (In German.) Springer, Berlin, 1997.

G. D'Angiolini, M. Fortini, M. Signore. SIDI: A Metainformation Management System for Survey Quality Control. Working Paper No. 13, UN-ECE/METIS, Work Session on Statistical Meta-data, 7 pp., 1996.

P.L. Darius et al. Modeling Metadata. *Statistical Journal of the United Nations Economic Commission for Europe* 10(2): 171-179, 1993.

M. Denk. Metadata Driven Production of Statistical Aggregates. Diploma Thesis, Institut für Statistik, Operations Research und Computerverfahren, Universität Wien, 1999.

EUROSTAT, editor. *Proc. Statistical Meta-Information Systems.* Office for Official Publications, Luxembourg, 1993.

Y. Franchet. Funktion der Statistik bei der europäischen Integration. (In German.) *Allgemeines Statistisches Archiv* 79(1): 18-25, 1995.

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

K.A. Froeschl. On Standards of Formal Communication in Statistics. Working Paper No. 16, UN-ECE/METIS, Work Session on Statistical Metadata, 12 pp., 1999.

D.W. Gillman and M.V. Appel. Metadata Database Development at the Census Bureau. Working Paper, UN/ECE Conference of European Statisticians, Work Session on Statistical Metadata (METIS), 1994.

D.W. Gillman, M.V. Appel, W.P. LaPlant, Jr. Design Principles for a Unified Statistical Data/Metadata System. In P. Svensson and J.C. French, editors, *Proc. Scientific and Statistical Database Management* (Proc. 8th SSDBM), 150-155, 1996.

D.W. Gillman, M.V. Appel, S.N. Highsmith, Jr. Building a Statistical Metadata Repository at the U.S. Bureau of the Census. Working Paper No. 11, UN/ECE Conference of European Statisticians, Work Session on Statistical Metadata (METIS), 18 pp., 1998.

R.B. Graves, F.E. Hutton, G. Deecker. Information Holdings Within Statistics Canada: A Framework. Working Paper, UN/ECE Conference of European Statisticians, Work Session on Statistical Metadata (METIS), 27 pp., 1993.

W. Grossmann. Metadata. In S. Kotz, C.B. Read, D.L. Banks (Eds.). *Encyclopedia of Statistical Sciences*, Update Volume 3, 811-815. Wiley & Sons, New York, 1999.

W. Grossmann et al. IDARESA Summary. Technical Progress Report 0.1/4. Wien, 21 pp., 1998.

D.J. Hand. Statistical Expert Systems. *Chance* 7(1): 28-34, 1994.

ISO. Reference Model for Data Management. ISO/IEC 10032, 1995(E).

W. Keller and W.F.H. Ypma. Electronic Data Interchange for Statistical Data Collection. In *New Techniques and Technologies for Statistics* (Proc. NTTS '95), 129-139. IOS Press, Amsterdam, 1997.

J.-P. Kent. Take Care of the Meta, and the Meta Will Take Care of the Data. Working Paper No. 6, UN/ECE Conference of European Statisticians, Work Session on Statistical Metadata (METIS), 1998.

J.-P. Kent and M. Schuerhoff. Some Thoughts About a Metadata Management System. In *Scientific and Statistical Database Management* (Proc. 9th SSDBM), 174-185, 1997.

W. Krug. Some Proposals for Enhancing Cooperation between Academic and Official Statisticians. *Allgemeines Statistisches Archiv* 82(3/4): 352-368, 1998.

METIS. Standards for Statistical Metadata on the Internet. Working Paper No. 2. UN/ECE Conference of European Statisticians, Work Session on Statistical Metadata (METIS), 11 pp., 1998.

METIS. Guidelines for Statistical Metadata on the Internet. Working Paper No. 23. UN/ECE Conference of European Statisticians, Work Session on Statistical Metadata (METIS), 4 pp., 1998.

OECD (G. Petit et al.). List of Metadata Items for OECD's Main Economic Indicators. Working Paper No. 25, UN/ECE Conference of European Statisticians, Work Session on Statistical Metadata (METIS), 25 pp., 1996.

J. Oleñski. Practical Problems of Implementing Metadata Standards in Official Statistics. In P. Svensson and J.C. French, editors, *Proc. Scientific and Statistical Database Management* (Proc. 8th SSDBM), 130-147, 1996.

M.M. Parker and R.J. Benson. *Information Economics–Linking Business Performance to Information Technology*. Prentice Hall, Englewood Cliffs, NJ, 1988.

T.M. Porter. Statistics, Social Science, and the Culture of Objectivity. *Österr. Zeitschrift für Geschichtswissenschaft* 7(2): 177-191, 1996.

D. Prazenska. Common Terminology of METIS. Working Paper No. 9, UN/ECE Conference of European Statisticians, Work Session on Statistical Metadata (METIS), 48 pp., 1996.

M. Rafanelli and F.L. Ricci. MEFISTO: A Functional Model for Statistical Entities. *IEEE Transactions on Knowledge and Data Engineering* 5(4): 670-681, 1993.

D. Ramprakash. European Plan for Research in Official Statistics. Draft Paper, 42 pp., 1998

H. Sato. A Data Model, Knowledge Base, and Natural Language Processing for Sharing a Large Statistical Database. In M. Rafanelli et al., editors, *Statistical and Scientific Database Management* (Proc. 4th SSDBM), 207-225. Springer (LNCS 339), Berlin, 1989.

M. Schuerhoff. *BLAISE as a Statistical Control Center.* Bulletin of the ISI, Proc. 49th Session (Firenze), Book 2, IN17.1, 273-282, 1993.

E. Schulte Nordholt. Imputation: Methods, Simulation Experiments and Practical Examples. *International Statistical Review* 66(2): 157-180, 1998.

A.P. Sheth and J.A. Larson. Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases. *ACM Computing Surveys* 22(3): 83-236, 1990.

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

B. Sundgren. Organizing the Metainformation Systems of a Statistical Office. Working Paper No. 3, UN/ECE Conference of European Statisticians, Work Session on Statistical Metadata (METIS), 66 pp., 1992.

B. Sundgren. Statistical Metainformation Systems–Pragmatics, Semantics, Syntactics. *Statistical Journal of the United Nations Economic Commission for Europe* 10(2): 121-142, 1993.

B. Sundgren. An Information Systems Architecture for National and International Statistical Organisations. WWW-published manuscript, http://www.imim.scb.se/, 54 pp., 1997.

A. Sutcliffe and U. Patel. Analysing Requirements for Internet Information Delivery. *Research in Official Statistics* 0: 83-100, 1998.

T.J. Teorey. *Database Modeling and Design–The Entity-Relationship Approach.* Morgan Kaufmann Publ., San Mateo, Ca., 1990.

G. van den Berg and E. de Feber. Definition and Use of Meta-Data in Statistical Data Processing. In H. Hinterberger and J.C. French, editors, *Statistical and Scientific Database Management* (Proc. 6th SSDBM), 290-306. ETH, Zürich, 1992.

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

G. Zettl. Zentrales Tabellenformat. Internal Report (in German). ÖSTAT, Wien, 52pp., 1997.

An excerpt of an earlier version of this paper was delivered orally to the annual convention of the Austrian Statistical Society, Linz, Upper Austria, April 15-17, 1998.

Acknowledgement

I gratefully acknowledge the numerous stimulating discussions on the topic I had with Wilfried Grossmann over the last couple of years. Thanks also to Michaela Denk, Günter Zettl, and the anonymous referees for pointing out various weak points in my arguing.

Author's address:

Dr. Karl Anton Froeschl
Institut für Statistik, Operations Research und Computerverfahren
Universität Wien
Universitätsstraße 5
A-1010 Wien

Tel. (+43 1) 42 77 - 38614
Fax (+43 1) 42 77 - 9386
E-Mail: Karl.Anton.Froeschl@UniVie.ac.at