Industrial Statistics Revisited: From Footnotes to Meta-Information Management

UNIDO's Integrated Statistical Data and Data Documentation Framework

Karl A. Froeschl^A, Tetsuo Yamada^{B*} and Roman Kudrna^A

A Department of Statistics and Decision Support Systems
University of Vienna

B Statistics and Information Networks Branch, UNIDO

Dedicated to Gerhart Bruckmann on the occasion of his 70th birthday

Abstract: UNIDO maintains a long tradition in the compilation of international statistics on industrial production and, particularly, has developed a suite of tools and techniques for improving third-party data, mainly supplied directly or indirectly by National Statistics Offices, to enable and enhance both cross-country and long-term comparability. However, changing IT environments, socio-economic conditions, and customer requirements increasingly challenge established procedures and behaviors. After reviewing (i) the relevance of industrial statistics in general and that of international industrial statistics in particular for the industrial development in the macro economic framework, (ii) the importance of data comparability on those statistics for accurate and objective analysis with regard to the industrial development, and (iii) UNIDO's specific efforts in increasing cross-country data comparability, this paper discusses a proposal for an integrated data and data documentation framework aimed at (i) recording all measures taken by UNIDO and (ii) reporting on all residual inconsistencies and deficiencies adversely affecting data comparability and interpretation. The proposal is illustrated by a prototype implementation using current UNIDO data.

* The designations employed and the presentation of material in this article do not imply the expression of any opinion whatsoever on the part of secretariat of the United Nations Industrial Development Organization (UNIDO). The opinions, figures and estimates set forth in signed articles are the responsibility of the authors, and should not necessarily be considered as reflecting the views or carry the endorsement of UNIDO.

1 Role, Relevance, and Data Sources of Industrial Statistics

One of the top agenda in the international community relates to poverty alleviation. According to recent development economics, economic growth is essential for poverty alleviation. The most recent macro-economic growth empirics (or, the new vision for such research) call for analysis at the sectoral and the sub-sectoral levels in order to identify the sources of growth in individual countries or in individual groups of similar countries and, consequently, to develop relevant policy recommendations.

Thus, in the era of globalization of economy, monitoring of sustainable (particularly in terms of economic development) industrial development and related economic-growth analysis require specialized and detailed economic industrial statistics as an indispensable information basis to assess implications of the globalization process for individual countries. This is because meaningful monitoring, interpretation and prediction of structure and performance relating to demand, supply, trade, investment, etc., can be made only empirically. Viable strategies, policies and programs for sustainable industrial development and investment cannot be formulated unless they are well prepared based on such statistics and analysis. Those that are not based on relevant empirical evidence would be subject to dispute.

Taking the above into consideration and following the inter-agency arrangements with regard to compilation and dissemination of international statistical information in the international statistical community, the United Nations Industrial Development Organization (UNIDO)¹, in collaboration with the Organization for Economic Co-operation and Development (OECD), assumes the sole responsibility for the collection, development, maintenance and dissemination of worldwide key industrial statistics at detailed sub-sectoral levels of the manufacturing sector for industrial production in individual countries. These international data on the key industrial statistics together with other available international databases on international commodity trade, national accounts, etc. enable researchers to carry out country-specific or cross-country empirical economic research, for instance, on

- manufacturing production and its growth, pattern and cross-country location,
- international trade in relation to manufacturing production and comparative advantage,
- production factors of manufacturing,
- manufacturing productivity, and
- structure (e.g., production structure, trade structure, employment structure, market structure) of the manufacturing sector.

The industrial statistics databases maintained by UNIDO, called as *INDSTAT Databases*, are arranged in accordance with the 3- and 4-digit levels of Revision 2 and

¹ As a Specialized Agency of the United Nations system, UNIDO, which is based in Vienna, Austria, deals with the industrial development in developing and transition-economy countries.

3, respectively, of the International Standard Industrial Classification of All Economic Activities (ISIC)².

UNIDO requests individual countries to provide available national data on the selected industrial statistics in accordance with either Revision 2 or Revision 3 of the ISIC system, depending on the country's current national industrial classification scheme in use³. Depending on the revision level of ISIC with which the reported data are arranged, data are stored in one of the three INDSTAT databases, INDSTAT3, INDSTAT4 and INDSTAT-Rev3, run in parallel⁴. Data conforming to ISIC (Rev.3) are entered in INDSTAT-Rev3 and at the same time, whenever possible, converted to Revision 2 compliant structures, that is conforming to the 3-digit level of Revision 2, and entered in INDSTAT3 as well. This is done mainly in favor of maintaining updated historical time series ranging back as far as 1963.

Data items in the INDSTAT Databases are limited to those for which the majority of developing countries compile their national data on the *regular* basis. Furthermore, the databases focus on a *minimum* range of economic industrial statistics from the viewpoint of reduced reporting burden on national statistical offices (NSOs)⁵. Say, for instance, formulation of a programme to secure food supply would require quantitative data (i.e., data in physical measuring units) regarding arable land, cultivated areas, demand-supply balance of food products, etc. Similarly, for the industrial development aiming at sustainable (here, it is referred to as economic sustainability only) economic development/growth in any country/region, a set of *key* economic industrial statistics is required as the basic statistical information source⁶.

The remaining sections of this paper are organized as follows. Section 2 investigates several sources of data incomparability and summarizes UNIDO's countermeasures towards achieving acceptable levels of data consistency. Section 3, after having discussed several insufficiencies and shortcomings of traditional data management practices, sketches – both formally and technically – the envisaged design of an integrated data/metadata system architecture closely linking industrial statistics data and data documentation. In Section 4, a prototype implementation of this design is presented, using real-data examples for illustration. Finally, Section 5 concludes with some remarks concerning implications, opportunities, and difficulties of the proposed statistical information management framework.

² UNIDO Industrial Statistics at the 3-digit Level of ISIC (Rev.2) (INDSTAT3); UNIDO Industrial Statistics at the 4-digit Level of ISIC (Rev.2) (INDSTAT4); and UNIDO Industrial Statistics at the 3- and 4-digit Levels of ISIC (Rev.3) (INDSTAT-Rev3). In addition to those INDSTAT databases, UNIDO also maintains several other databases at sectoral and sub-sectoral levels of the manufacturing sector.

³ Presently, approximately two-third of some 120 regularly-reporting countries and areas are in a position to report their data in Revision 3.

⁴ Strictly speaking, INDSTAT Databases correspond to a *file system* hosting time series sharing, however, a coordinated set of classification criteria and storage formats.

⁵ The INDSTAT Databases refer to annual time-series on the following statistics for individual countries and areas: Number of establishments; Total employment; Female employment; Wages and salaries paid to employees; Output; Value added; Gross fixed capital formation; and Production indexes (INDSTAT3 only).

⁶ It should be noted that commodity-level statistics (e.g., for production, consumption, trade) *in physical units* are limited in terms of commodity coverage. Furthermore, they are not really economic statistics and, thus, of limited validity in decision making for viable economic development and investment.

2 Quality and Consistency of Industrial Statistics Data

In the context of globalization, international statistics for many social, economic and environmental areas that are collected, maintained and disseminated by international data sources have been increasingly demanded not only for country-specific analysis but also for cross-country analysis. For consistent country analysis on industry in relation to other countries, cross-country comparable data on those statistics are required. To secure cross-country comparability of individual data that are stored in the INDSTAT Databases, UNIDO carries out thorough quality check and adjustment of the data reported by national statistical offices on the basis of ISIC as well as international standard concepts and definitions.

The industrial statistics under the current consideration are highly detailed in terms of industry classification. Furthermore, they are measured/valued according to varying concepts and definitions. Thus, their information value is high but, at the same time, their reported data tend to be incomplete and inconsistent in terms of comparability across countries and sometimes even across variables, industries and over time because of possible differences in employed national standards, methods, and survey scope. Therefore, the validity of each reported datum is judged carefully on the basis of international standards.

Quality of statistical data has several dimensions. They include, in particular, *relevance*, *accuracy*, *timeliness*, *accessibility*, *comparability*, *coherence* and *completeness*. UNIDO's quality-assurance work on reported national data for those quality dimensions refers to, whenever possible, (i) identification and documentation of the sources of existing deviation from international standards and recommendations based on the respective metadata given by NSOs, (ii) adjustment of incomparable/incoherent data by employing available supplementary information, (iii) imputation and estimation of missing data by use of available supplementary information and econometric methods, and (iv) improvement of the organization's data publication and dissemination services together with the development of user-friendly interfaces for data manipulation incorporated in its data dissemination products.

2.1 Potential Sources of Cross-country Incomparability of Data

Major determinants for the extent of international comparability of reported data are data coverage (or scope of the national establishment survey), employed concept and definition of the variable in question, and the national industrial classification, which are discussed below in some detail.

Data coverage:

Often, reported data are known to exclude a significant portion of industrial activity, either because the coverage of small-scale establishments may be incomplete (e.g., exclusion of the establishments below a certain survey cut-off-point in terms of their size), the data may refer only to a certain area of the county (e.g., urban area, major

provinces), or they may refer to only a part of the manufacturing sector (e.g., formal sector, registered establishments, selected industries, etc.). This characteristic is certainly the most challenging of all sources of data incomparability because adjusting for coverage involves the attempt to quantify what is not there. The problem of data coverage may be broken down into three parts: (i) incomplete or varying degrees of coverage of establishments; (ii) non-reporting of data by surveyed establishments, and (iii) the failure to adjust for non-response.

While the question of the treatment of non-response is basic for the data user, it has not received the attention that it deserves among many national data producers. Some countries adjust their data for non-response, and others do not. The latter usually provide some crude measure of the extent of non-response, which can be used to assess the quality of the data. However, some countries fail to address the question altogether. The International Recommendations (UN, 1983) specifically request such information, and perhaps this is an area where improvements in national reporting practices may be anticipated.

Concepts and definitions:

Differences in concept or definition are variable-specific although their numerical effects may vary across industries (i.e., ISIC categories). In reporting their industrial data, most countries conform to the United Nations' recommendations. Even among those countries that do not, the international standards provide a convenient reference point for comparing all variations in national reporting practices.

Employment: For the majority of the countries, employment data refer to number of employees. However, in some cases data refer to number of persons engaged. For a few countries, the definition changes over time. In general, no supplementary information for standardization of reported employment data is available. Any use of employment data, therefore, requires caution, particularly in those cases where definition changes over time.

Wages and salaries: In the reporting of wages and salaries, the most common differences between national practices and the international recommendations relate to the inclusion of payments to family workers and of employers' contributions to social security schemes or the exclusion of payments-in-kind. The numerical effects of these differences, although not known, are probably of small consequence both within and between countries, compared to the effects of differences in survey cut-off point.

Output and value added: Among the variations in concept that may apply to the data on output and value added, the most important (in terms of measuring consequence to data) are: (i) the difference between the national accounting concept and the industrial census concept; and (ii) the difference between valuation at producers' prices, valuation in factor values and other valuations. The main difference between the national accounting concept and the industrial census concept is in the treatment of non-industrial services. On the other hand, that in the different valuations is in the treatment of indirect taxes and subsidies. These differences can be significant, and should be taken into account especially if comparisons between countries are being made.

In summary, cross-country incomparability comes from:

- incomplete coverage of the variables;
- incomplete time series and data inconsistency over time;
- a lack of relevant metadata that indicate the comparability and quality of the reported statistical data, or
- cross-country, cross-variable and/or cross-year incomparability due to different/ changing combinations of ISIC categories, possibly missing adjustment of non-response, different definitions of variables, etc.

All these problems limit the usefulness of the data in cross-country comparative analysis, country-specific trend analysis, etc. of industrial growth and structure. In principle, improvement of data quality can be done only by the responsible NSOs.

2.2 Improving Data Consistency

To improve the international usefulness of the data reported by NSOs, not only the correction of simple reporting errors based on the NSOs' responses to concerned data queries by UNIDO but also adjustment of the reported data to achieve international standards as well as estimation and imputation of missing data to increase data coverage are required. For UNIDO, these tasks are possible only if appropriate supplementary statistical information is available. Furthermore, they require sector-specific statistical as well as economic knowledge.

UNIDO statisticians take considerable care in ensuring data consistency in the process of enlarging the INDSTAT databases and in improving the international comparability of their contents. However, due to inconsistency inherent to many series reported by primary and other sources, it is felt that a *final screening* of the data is needed. The purpose of this final screening is to diagnose and display 'abnormal' entries in the databases, to allow for possible corrections. The final screening takes place in two phases. First, possible abnormalities are identified through a computerized procedure. Secondly, UNIDO statisticians redress, to the extent possible, the identified abnormalities.

UNIDO's procedures to treat the national data reported by each NSO are summarized as the following:

- a) Pre-filling of the out-going *UNIDO General Industrial Statistics Questionnaire* with previously reported statistical and metadata for their possible revision by the NSO.
- b) Upon receipt of the questionnaire completed and returned to UNIDO by the NSO, (i) manual detection and, if possible, correction of incoherent data (e.g., mismatch between sum of components or disaggregated data and given total) and dubious data (e.g., obvious typo errors, abnormally fluctuating data, those resulting abnormal relations between variables), (ii) adjustment of reported statistics to desired statistics (e.g., if data on wage rate are reported instead of wage bill, deriving data on wage bill from reported data on per employee wage rate and those on the number of employees), and (iii) if appropriate, re-description of the provided metadata from the viewpoint of international comparability.

- c) If the validity of the dubious data cannot be judged, enquiry to the NSO for clarification of them, and correction of them in accordance with the clarification made by the NSO.
- d) Incorporation of the corrected and adjusted data in a relevant INDSTAT database. If the reported data are in Revision 3 of ISIC, incorporation of the data in INDSTAT-Rev3 database and, at the same time, conversion of them to those at the 3-digit level of ISIC Revision 2 and incorporation of the converted data in INDSTAT3 database. Available metadata are incorporated in a *metadata* file common to all INDSTAT databases.
- e) Computer-based screening (consistency and coherency check) of the initially updated INDSTAT databases to further detecting dubious data, and manual correction/rejection of the detected data (if deemed necessary, clarification is requested to the concerned NSO).
- f) Computer-based estimation of missing data by estimating inter-variable relations as country-specific time-series regression equations (such estimation is attempted only for four variables, namely total employment, wage bill, output and value added).

Achievement of the complete data coverage and, in particular, that of the cross-country comparability of data by employing common standards across countries are the top priorities in the maintenance of the INDSTAT Databases. However, such achievement by UNIDO alone is limited without consensus in definition, adoption and implementation of international standards in the international statistical community involving all NSOs.

2.3 User-oriented Metadata in Support of INDSTAT Databases

Cross-country incomparability, in many cases, of reported data is due to different national needs and practices in industrial statistics for legal, economic and policy purposes. However, adjustment of incomparable data in line with international comparability as well as estimation of missing data is often difficult to be made. Thus, the coverage and the extent of the comparability of the available data in each INDSTAT database are limited. To assure sound use of those data, UNIDO, as in the case of any other supplier of statistical information, is obligated to provide users with, in addition to general information for the overall database, relevant and detailed qualitative information (e.g., information concerning the deviations from international standards or norms) that indicates the applicability/limitation of individual disseminated statistical figures in terms of international comparability.

Preparation of appropriate statistical background information (metadata) in support of INDSTAT Databases requires concrete and well-documented metadata inputs from primary data compilers. Thus, UNIDO requests NSOs to provide, together with available statistical data, such *descriptive* information through its industrial statistics country-questionnaire. The key items for which the organization needs to obtain meta-information include:

- name of the supplier of the statistical data (i.e., the reporting NSO),
- basic source of data (e.g., annual industrial survey),
- major deviations from ISIC,

- reference period (e.g., calendar year),
- scope of the annual survey,
- method of enumeration and data compilation,
- treatment of non-response,
- concepts and definitions of variables,
- related national statistical publications, and
- cell-level footnotes, if any.

The metadata that are provided by NSOs often do not explicitly indicate deviations from international standards. In such cases, UNIDO attempts to re-describe/re-arrange the provided metadata into explicit information concerning the deviations from the international standards. This is often a difficult task and requires additional clarifications from the concerned NSO.

Countries have been increasingly introducing international statistical standards because of the reasons stated earlier, which has resulted in less needs of metadata on international statistics and in more relevant metadata inputs from NSOs from the viewpoint of international standards in recent years.

As briefly stated above, UNIDO makes considerable efforts to adjust the official data reported by NSOs to increase the international comparability of those data and to estimate missing observations to increase the data availability of its databases. The data adjustments are made, as much as possible, in a way that the adjusted data would be coherent with the corresponding reported official data in order to avoid the situations of conflicting data between the two data sources and consequent confusion among users. Furthermore, adjusted data and estimates generated by UNIDO are *footnoted* as UNIDO estimates.

3 Design of an Industrial Statistics Information System

Much like other institutions collecting, analyzing, and re-distributing statistical information, UNIDO increasingly faces a necessity of adapting its statistical data management procedures in response to changes in the surrounding informational environment. A particular drawback in the traditional way of managing statistical *background* information is the apparent disintegration of this background and the data itself which, to a great deal, used to be bridged by staff intervention and expertise. While, in former days, it has been a real possibility to keep in touch with both data suppliers and data users to exchange and communicate important background knowledge about the data conveyed, especially in modern networked information dissemination schemes these ancestral procedures fall short of emerging demands and, hence, need to be replaced with something more powerful, flexible, and sustainable.

First of all, any tangible improvement rests on a tight integration of both data and data documentation in a single coordinated information system providing a

homogeneous environment for both, data processing and retrieval,⁷ as such an integrated approach is the obvious precondition to the consistent retrieval of *all* additional information – any type of annotated information including especially table footnotes – attached to accessed statistics data without specific, “discontinued” inquiry measures or actions. More specifically, since statistical background information on industrial statistics ranges, as already pointed out in Subsection 2.3, from simple footnotes to single table cells (or time series entries) to definitions or methodological remarks, the *uniform* and formalized dealing with non-numerical information becomes a mandatory prerequisite of system design. To date, the lack of a systematic management of the variegated sorts of statistical meta-information, along with disintegrated storage of data on the one side and (mostly text-based) comments on the data on the other, has precluded any reasonable solution for an efficient, comprehensive and flexible statistical information service open to both internal and external user classes.

Figure 1. Data Item–Annotation Association

A major (conceptual) effort to re-build UNIDO’s INDSTAT Databases was launched in late 1999 with its primary aim to exploit latest metadata management ideas and technologies in making information available in a coherent way yet circumventing, as far as attainable, manual output preparation. Altogether, system design was to strictly follow a couple of methodological principles such as:

- Data documentation is conceived of as decomposed into single statements, henceforth called ‘annotations’, each belonging to a particular (taxonomic) documentation *class*.

⁷ INDSTAT’s data retrieval system generates various outputs including the *International Yearbook of Industrial Statistics* (annual hardcopy publication), country statistical briefs (daily hardcopy publication), statistical country tables (website), sales versions of INDSTAT (CD-ROM), and pre-filled *General Industrial Statistics Questionnaire* (both hard copy and MS-Excel format). Currently, data users’ access to available country-specific metadata is limited to the *Yearbook*.

- Each statistical datum (“data item”) might give rise to documentary entries of one or more documentation classes; hence, a relational association (more specifically, a *bipartite graph*) structure between statistical data and annotative information has to be made provision for (cf. Figure 1).
- Annotative information could refer to more or less extended (regular) collections of statistical data, thus also suggesting a parsimonious mechanism of annotation assignment to statistical data. In particular in view of the *dimensional* structure of INDSTAT data, documentary information relates to dimensional structures as well.
- Every access to statistical data must entail the access to attached annotations as far as they are actually affected (yet, conversely, access to data documentation does not necessarily imply also access to statistics data). Annotation access, of course, is therefore controlled by the structure of annotation assignment.
- It should be possible to feed back the results of statistical procedures (like imputations and estimations) applied to INDSTAT data right into the system in a consistent way.
- Data maintenance, particularly measures taken by UNIDO for the sake of improvement of data coverage and consistency (cf. steps b)–f) mentioned in Subsection 2.2) ought to be routinely tracked; in other words, the system should behave “auto-documentary” to the highest degree possible.
- Access to and maintenance of both data and documentation ought to be as simple and uniform as possible.

In summing up, the most significant features of system design regard (i) the formal coordination of the data item—annotation assignment and (ii) the storage of data documentation by way of its deconstruction into annotations (or documentation “chunks” such as single definitions, footnotes, etc.) broken down in terms of a pre-defined annotation classification. Moreover, the design strives for generality in that, despite its specific origin, fundamental characteristics are independent of the INDSTAT application domain. Technically, components of the proposed approach revolve around a “dual” data/metadata holding structure sharing a unified conceptual storage model as well as an easy-to-use retrieval interface allowing to state information interests – “queries” – in *descriptive* (rather than procedural) terms.

Regardless of the shortcomings mentioned, data preparation procedures at UNIDO lend themselves favorably to system re-design in several ways. First, traditional INDSTAT data organization has ISIC-based time series stored, by and large, in a multi-dimensional scheme admitting easy reference by economic indicator, time (year), ISIC code, and country, respectively. Moreover, industrial statistics data production background information is collected pre-structured according to a kind of template to be filled in by each contributing country. These “country notes”, although weakly formatted, actually prepare the field for annotation classification. Additionally, data item-based footnotes are already maintained in a separate footnote file indexed much the same way ISIC-data itself is (e.g., by country, year, indicator, and ISIC category). Finally, UNIDO has arranged for a stage model of recording the “life-cycle” of a data item from originally reported values through several well-discerned processing states, or *stages*, any such data item might undergo within UNIDO: while the first stage always contains the data as officially reported by national statistical offices (NSOs, or OECD); later stages – by incorporating further information of decreasingly authoritative sources

or methods – cumulate the effects of gap filling, data adjustment and model-based estimates.

As mentioned previously, INDSTAT data, first of all, refers to either of the ISIC Revisions in use. ISIC is a multi-level classification system and, as such, is giving rise to an often occurring nuisance: a good many countries, for one reason or another, in their recurrent data reporting object to fully concord to the whole set of ISIC categories (of the respective ISIC Revision used), instead collapsing or re-allocating groups of categories according to their own preferences⁸. Evidently, this practice defeats straightforward tabular alignment of different countries' time series and, thus, runs counter to making country data comparable. However, from a quality point of view, it is still preferable to have countries making the imperfections of their data explicit rather than letting such deviations go unnoticed at all. Despite the headache caused by classification irregularities of reported data and the obvious need of any practical information system for addressing this issue, in what follows little attention will be paid to this peculiarity.

3.1 Formal Meta-information System Structure

Little surprise, overall system design starts with the notion of a data cube (now well-known from the field of *data warehousing*; e.g. cf. Kimball, 1996). Such a data cube resembles a multi-dimensional (cross-sectional) statistical table with cells each holding the value of some indicator (aggregate value, statistical datum) broken down with respect to a couple of cross-classifications (table dimensions). In the present context, though, the “concept” of the data cube is generalized significantly in two ways. First, cross-classifications are used as a *formal* device for any kind of data segmentation including dimensions for spatial and temporal breakdown as well as dimensions for separating different data processing stages and even different types of indicators. Clearly, each such dimension – named *cube edge*, henceforth – entertains its particular semantics and must thus be treated differently especially from the processing point of view. In a sense, this “enlarged” data cube might be conceived simply as a peculiar kind of higher-dimensional spreadsheet where to fill in all stored data. Secondly, cube cell content distinguishes between statistical and *annotation* data, letting ‘annotation’ denote any kind of remark or piece of documentation directly associated with a data cube cell. To allow for a subject-matter subdivision of annotations, the formal cross-classification concept symmetrically extends to specific “annotative cube edges” each breaking down documentation into different annotation categories (cf. Silver, 1993).

The proposed information system architecture (Froeschl and Yamada, 2000) in fact comprises two such cubes, one for statistical data (the *data cube* proper) and another for annotations (the *annotation cube*), interrelated by a set of shared cube edges. More specifically, let $d_1, \dots, d_p$ denote the set of $p > 0$ formal edges of the data cube; let further $B_i \equiv B(d_i)$ denote the (extensional) modality sets associated with edge d_i ($1 \leq i \leq p$). On top of these basic set extensions, grouping functions $\gamma_i : C_i \rightarrow 2^{B_i}$ are declared, mapping the categories introduced, depending on the semantics of dimension

⁸ ISIC category combinations are, for some cases at least, not avoidable due to a mismatch between the national classification and ISIC or statistical confidentiality rules.

i , to non-empty subsets of B_i such that, in all cases, $B_i \subseteq C_i$, that is, in particular, all modalities of B_i re-appear in C_i . Now, each p -tuple $\mathbf{x} \in \times_{i=1}^p C_i$ – or “crossing” – identifies a (possible) statistical datum of the data cube, the value of which is referred to as $\delta(\mathbf{x})$; in database terms, $\mathbf{x}$ acts as an access key. Apparently, the value type of $\delta(\mathbf{x})$ depends particularly on the statistical indicator referenced by $\mathbf{x}$.

Depending on edge (and, particularly, indicator) semantics, various integrity constraints might be imposed on related $\mathbf{x}$ -values. Naturally, the *data space* (cf. Rafanelli and Ricci, 1993) $\mathbf{D} \equiv \times_{i=1}^p C_i$ hosts a partial ordering on categories $\mathbf{x}' \prec \mathbf{x}''$, $\mathbf{x}', \mathbf{x}'' \in \mathbf{D}$, established in terms of grouping functions, that is $\mathbf{x}' \prec \mathbf{x}'' \Leftrightarrow \gamma_i(x'_i) \subseteq \gamma_i(x''_i)$, $1 \leq i \leq p$ and $\gamma_j(x'_j) \subset \gamma_j(x''_j)$ for some $j \in \{1, \dots, p\}$. Now, for reasons of consistency, in case of a summarizable indicator, for $\mathbf{x}' \prec \mathbf{x}''$ it must necessarily hold that $\delta(\mathbf{x}') \leq \delta(\mathbf{x}'')$. For strictly additive indicators, storage redundancy can be avoided by populating only those cells of a data cube not obtainable from already stored values by summation.

With respect to the annotation cube, the set-up is entirely analogous except that, normally, at least one further “annotative” edge is added. Thus, an annotation cube comprises the dimensions $d_1, \dots, d_p, d_{p+1}, \dots, d_q$ for $q \geq 0$, giving rise to its *annotation space* $\mathbf{A} \equiv \times_{i=1}^q C_i$ with B_j , C_j , and γ_j ($p+1 \leq j \leq q$) defined accordingly. Likewise, each q -tuple $\mathbf{y} \in \times_{i=1}^q C_i$ identifies an individual cell of the annotation cube, the value of which is referred to as $\alpha(\mathbf{y})$. Note that, by construction, $\alpha(\mathbf{y})$ pools *all* documentary information relating to $\mathbf{y}$, implying that a finer-grained representation would call for the introduction of additional annotation categories (that is, an expansion of some B_j , $p+1 \leq j \leq q$) or even of a whole new independent annotative edge itself.

Quite in contrast to the data cube, the implication $\alpha(\mathbf{y}') \Rightarrow \alpha(\mathbf{y}'')$ holds whenever $\mathbf{y}'' \prec \mathbf{y}'$ for $\mathbf{y}', \mathbf{y}'' \in \mathbf{A}$. In other words, it is sufficient to store “maximal” annotations only; $\mathbf{y}'$ is termed the *more general* (than $\mathbf{y}''$) annotation, $\mathbf{y}''$ the *more specific* (than $\mathbf{y}'$) annotation.

By design, data and annotation cubes are interrelated easily through defining a *projection* of elements $\mathbf{y} \in \mathbf{A}$ onto $\mathbf{D}$, viz. $\pi(\mathbf{y}) \equiv \pi(y_1, \dots, y_q) = (y_1, \dots, y_p) \in \mathbf{D}$. Correspondingly, some annotation $\mathbf{y} \in \mathbf{A}$ is “of relevance” for a datum $\mathbf{x} \in \mathbf{D}$ if and only if $\pi(\mathbf{y})$ and $\mathbf{x}$ intersect, that is $\gamma_i(\pi(\mathbf{y})_i) \cap \gamma_i(x_i) \neq \emptyset$ for $1 \leq i \leq p$.

Given this twin cube system design, query processing becomes quite a straight matter. Skipping subtle details, a pure documentation query $\Omega_A = \{\omega_1, \dots, \omega_r\} \subseteq \mathbf{A}$, $r \geq 1$, amounts to checking whether there are – not necessarily distinct! – annotations $\alpha(\mathbf{y}_1), \dots, \alpha(\mathbf{y}_r)$ stored (that is, non-empty cube cells) for which either $\omega_k = \mathbf{y}_k$ or $\omega_k \prec \mathbf{y}_k$, $\mathbf{y}_k \in \mathbf{A}$ ($1 \leq k \leq r$). Owing to summarizability conditions, data queries might turn slightly more complex (cf. Chen et al., 1989). In particular, assuming strict additivity, for any disjoint union yielding $\mathbf{x} \in \mathbf{D}$, say $\mathbf{x} = \mathbf{x}' \oplus \mathbf{x}''$, $\delta(\mathbf{x}' \oplus \mathbf{x}'') = \delta(\mathbf{x}') + \delta(\mathbf{x}'')$. Hence, generating a response to queried $\omega \in \Omega_D \subseteq \mathbf{D}$

typically amounts to find a disjoint subdivision $\{\mathbf{x}_1, \dots, \mathbf{x}_u\} \subseteq \mathbf{D}$, subject to simple “tiling” conditions, such that (i) all $\delta(\mathbf{x}_k)$, $1 \leq k \leq u$, exist in the database, and (ii) $\mathbf{x}_1 \oplus \dots \oplus \mathbf{x}_u = \varpi \in \mathbf{D}$ with ϖ as “close” as possible to the queried ω (it might be decided that $\varpi \prec \omega$ all the time). For $u > 1$, the response value simply evaluates to $\sum_{k=1}^u \delta(\mathbf{x}_k)$; however, since determining ϖ leads to a combinatorial (Pareto-) optimization problem in general, heuristic shortcuts and/or user preferences (elicited interactively) have to be resorted to practically.

A queried $\omega \in \Omega_{\mathbf{D}}$ gives rise to a set-valued documentation query $\Omega_{\mathbf{A}} = \bigcup_{k=1}^u Y(\mathbf{x}_k)$ where $Y(\mathbf{x}) = \{\mathbf{y} \in \mathbf{A} \mid \pi(\mathbf{y}) = \mathbf{x}\}$. Contrary to pure documentation queries, in data queries no further constraints must be imposed on the induced $\Omega_{\mathbf{A}}$ with respect to dimensions $p+1, \dots, q$ (since this would cut off pertinent documentation elements). Apparently, in case that $\varpi \neq \omega$ there is no point in retrieving annotations to parts of ω actually not covered by the generated response ϖ (thus, $\Omega_{\mathbf{D}} = \{\omega\}$ and $\Omega_{\mathbf{A}} = \{Y(\omega)\}$ could in fact produce different documentary responses!).

This abridged account (for reasons of brevity) of system structure omits, of course, detailed design decisions practically determining system functionality. In particular, the essential structural aspects – such as classification hierarchies – of the grouping functions γ_i need to be clarified as these heavily influence the actual generation (and its complexity!) of responses to given queries, $\Omega_{\mathbf{D}}$, under general conditions. The following subsection sketches a possible roadmap of applying the devised framework to the INDSTAT domain.

3.2 Moving the Framework to the INDSTAT Context

As a first approximation, the formal framework outlined might assume the following definite shape in the context of INDSTAT. The data cube is composed out of five edges representing, in turn,

- a temporal breakdown dimension (‘time’, in years);
- a geographic breakdown dimension (‘geo’, countries);
- a breakdown of data in terms of ISIC (both Rev. 2 and 3, ‘industry’);
- a formal breakdown of data according to processing stages (‘stage’);
- another formal breakdown of data distinguishing between the (14) economic statistical indicators maintained within INDSTAT (‘indicator’).

While these dimensions conform to established INDSTAT practice, it is advisable to add a further *chronicle* edge (‘level’) used for keeping track of the sequence of values obtained for any data item determined in the data cube by a combination of the five mentioned cube edges. Technically, newly arriving values for a data item are enumerated by assigning them to (automatically) increased level numbers. So, by definition, INDSTAT data as reported by countries (or OECD) always enter the system at level 1.

Of the five remaining dimensions, only ‘industry’ gives rise to a grouping function defining a hierarchical layering of ISIC categories. As with any classification tree, first

the modality set B_{ISIC} is identified with all lowest-level categories comprised by the classification (3- or 4-digit levels, respectively). Then, for each grouping level in the hierarchy, the values occurring are associated with the corresponding elements of B_{ISIC} (actually, the ensuing set C_{ISIC} is further augmented with “artificial” codes redressing national coding mismatches; see above). Because of the linear ordering of chronicle levels, there never arises a need of value groupings for this cube edge (that is, substantially, C_{level} and B_{level} coincide). Likewise, none of the remaining cube edges uses value groupings at present.

Figure 2. Basic Twin Cube Model Structure

To the six cube edges introduced, the annotation cube adds a seventh, conveniently called *note class*, again used without any value groupings (that is, a flat list of annotation categories is deemed sufficient for the time being). Figure 2 highlights, in a simple block diagram, the relationship between data and annotation cubes, respectively, separating shared from non-shared cube edges. Practically, all modality domains of the cube edges in connection with grouping hierarchies (in case of ISIC Revisions) have to be stored as part of the model configuration. An excerpt for modeling cube edge-related information is shown in Figure 3.

According to Figure 3, each cube edge is implemented as a ‘Dimension’ object to which modalities (‘Code’ objects) are assigned. Each code may either denote a group of values (that is, other codes), or belong to any number of value groupings as group element, as controlled by the ‘CodeStatus’ field). Moreover, the data model formally distinguishes ‘Code’ from ‘CodeLabel’, this way allowing for storage of textual labels (e.g., prepared in different natural languages) separated from structural information. Figure 4 exhibits a small excerpt of ISIC code storage.

In practical data modeling, the multi-dimensional data cube structure is decomposed (mainly for efficiency reasons) into a ‘Master’ and a ‘Level’ segment as shown in Figure 5. The first segment carries the cube edges substantively important in primary data item selection while the second gathers actual data values further indexed by ‘stage’ and ‘level’ categories.

Figure 3. Cube Edge Data Model (excerpt)

DimValue	Dim	CodeSystem	Code	PrintLabel
53121	industry	ISIC2	3311	Sawmills, planing and other wood mills
53122	industry	ISIC3	3311	Medical, surgical and orthopaedic equipment

Figure 4. Code Storage Sample (simplified)

Figure 5. Data Cube Storage Model

This segmented design accounts specifically for inserting “updated” values into the database in that new ‘Level’ objects are simply added to the respective ‘Master’ object denoting the data item updated. Contrary to traditional INDSTAT practice, stacking update values of data items in terms of a ‘stage’ vector according to the processing scheme outlined in Subsection 2.3, this revised design allows *multiple* ‘stage’ values for each data item distinguished by (automatically assigned) level numbers. From a technical point of view, by means of stage/level-indexing the 6-dimensional data cube might be considered as a kind of multi-dimensional *spreadsheet* where, except for the level-1 layer, other cells are filled by update procedures, in particular by applying formulae or statistical (estimation/imputation) algorithms taking other cell contents as primary input, if need be. While the ‘level’-dimension discerns the different values computed, the ‘stage’-dimension records update-specific meta-information such as the statistical method applied, or the measure taken, by creating corresponding note entries in the annotation cube. Of course, in general, annotations originate from manual insertion most of the time, as large parts of the data documentation are provided either by NSOs or OECD submitting country data, or by UNIDO statisticians.

Figure 6. Annotation Cube Storage Model

The annotation cube also makes use of a segmented design, this time, however, separating actual note entries from the 7-dimensional cube structure as shown in Figure 6. The annotation representation scheme is in fact fairly parsimonious because (i) the segmented design facilitates multiple assignments of a note to cube cells, and (ii) the coverage of an assigned note could extend to any rectangular subsection of the annotation cube subsumed by an identifiable cube crossing (cf. Subsection 3.1). Different ensuing annotation assignment scenarios are depicted in Figure 7.

Annotation extensibility is controlled by the ‘XFlag’ field; if set, the applicability range of the assigned note includes all cube cells (data items, respectively) hierarchically subsumed as defined by the grouping structure. For instance, a note assigned, say, to ISIC code ‘15 – Food and Beverages’ with its XFlag set, applies to codes ‘151’ and ‘1511’ as well (cf. Figure 8), so, in case that either of these categories subsumed to code ‘15’ is queried, the note is detected and included in the response to the query. If, in particular, a group covers all modalities of a cube edge, the assigned annotation applies to the whole cube ‘slice’. In order to illustrate this feature, think of a peculiar concept definition used in a single country (or year, or country/year-combination, etc.); this, then, needs to be stated by a single note assignment only in the

annotation database since, further on, any single data item relating to this country (or year, or country/year-combination, etc.) is “inheriting” this note and can thus be retrieved in connection with any data item pertaining to this country (or year, or country/year-combination, etc.).

Figure 7. Different Annotation Assignment Options

Figure 8. Exploiting Code Hierarchies in Annotation Retrieval

Drawing together, as its main components system design comprises (i) the twin cube-data model surrounded by (ii) various auxiliary data structures capturing, essentially, the twin cube-schema. In partly encoding its own “data model”, system design is prepared to purposively harness this meta-information as *active metadata* in query processing. While, of course, introducing a bit of an overhead in storage and metadata management, the framework provides a neat and concise self-documentation useful also in providing both data and pertinent documentation to downstream systems and interfaces.

4 The UNIDIS Software Prototype Implementation

In 2001, a proof-of-concept software prototype has been implemented (using Microsoft SQL-Server and Windows-based client software platforms) in order to demonstrate the practical virtues of the proposed INDSTAT re-design. While using the whole scope of original INDSTAT data along with footnotes⁹ and (an excerpt of) definition and country notes documentation, the prototype was deliberately confined in its query processing functionality. Essentially, the prototype is to highlight (i) integrated annotation retrieval and (ii) self-annotative database updating as part of UNIDO’s data quality improvement policy. In what follows, only the client interface of the prototype is referred to.

Upon invocation, the UNIDIS client exhibits its main window as reproduced in Figure 9, with its ‘Query composer’ pane opened. This pane is used to state a (data) query by selecting a range of data items in terms of choosing (i) the indicator of interest, (ii) a set of ISIC categories relative to the selected ISIC Revision, and (iii) a range of years for (iv) some countries.

Before actually submitting the query to the underlying database system, a *query strategy* needs to be specified. Basically, this query strategy controls the choice of a particular stage/level combination for the data items selected. In the example shown in Figure 9, ‘latest Entry’ means that, regardless of the stage the value belongs to, always the latest available value of a data item is chosen (that is, the value inserted last into the data cube for the respective data item). Other query strategies allow selecting the very first (that is, level 1) value of data items, or restrict the query to particular stage categories.

After a query is submitted, it is processed according to the chosen query strategy, returning the numerical output in a standard tabular format on the ‘Result viewer’ pane of the UNIDIS client as shown in Figure 10. Due to its limited intelligence, the prototype would fail to compute the sum totals for requested ISIC category ‘311’ even if, for all requested 4-digit level codes, data were available in the database.

Note that, if data for more than one country are queried, the presentation appears layered country by country. Clearly, this simplified pane design is questionable especially for cross-country comparison whence, in realistic software, more flexible pivoting of presented tables is actually called for.

⁹ It must be noted, however, that most of the original INDSTAT footnotes simply disappear because of the different dealing with collapsed ISIC categories.

Figure 9. The UNIDIS Client Interface

If, in this presentation, one now was interested in additional annotative information for the cell corresponding to the category combination ‘Austria : ISIC 3112 : 1994’ by clicking on this cell¹⁰ another pane as depicted in Figure 11 would open up. For the sake of illustration, this annotation pane includes artificial notes to highlight the use of icons for indicating whether the note is directly attached to the data item, or rather inherited from a higher-level item by way of expansion. At present, only internal table cells are “click-able” for annotative information; however, in principle, also table stub and heading could be configured this way.

In order to better visualize INDSTAT data, the UNIDIS client offers an alternative mode of query response inspection by time series *charts* using the ‘Time series’ pane. To this end, a selection of rows is marked in the displayed table on the ‘Result viewer’ pane (see Figure 10); after switching to the ‘Time series’ pane the selected time series are displayed as illustrated in Figure 12. Each time series is assigned (internally) a different color in order to help discern the charts; annual data values are simply connected by colored straight lines.

¹⁰ A – simple to amend – weakness of the prototype is its inability to highlight whether or not a cell in fact carries annotative information.

UNIDIS Client

Query composer | **Result viewer** | Time series

Results for Indicator Number of employees Export...

Austria

ISIC 2	1988	1989	1990	1991	1992	1993	1994	1995
311	.	.	.	.	.	.	.	.
3111	14000	13700	14200	14000	14000	14408	13907	20185
3112	8500	8400	7800	7000	7000	5838	5401	6420
3113	1500	1500	1500	2000	2000	1479	1438	1773
3114	.	.	.	.	.	.	.	.
3115	.	.	.	.	.	.	.	.
3116	1300	1200	1200	1000	1000	990	912	1045
3117	11400	11600	12900	13000	13000	12975	13450	20590
3118	1300	1300	1400	1000	1000	1116	1067	2028
3119	3100	2900	3300	3000	3000	3078	2950	4121
3121	10500	10600	10500	10000	10000	10071	.	.
3122	.	.	.	.	.	.	.	.

Figure 10. UNIDIS Client Query Response Presentation

Figure 11. UNIDIS Client Annotation Pane

Figure 12. UNIDIS Client Graphical Output Presentation

Every now and then, INDSTAT time series are incomplete. In such cases, imputation of missing entries is a natural choice. Far from being statistically satisfying, the UNIDIS client explores a possible approach towards interactive “gap filling” such that a meticulous bookkeeping of all updates fed back to the database is warranted.

Figure 13. UNIDIS Prediction Assistant Parameter Summary

In order to estimate, for the sake of imputation or prediction, the annual value of some time series, the UNIDIS 'Prediction Assistant' is activated (by clicking on the 'Prediction...' control of the 'Time series' pane). As a first step, in this simplified set-up, one particular time series has to be chosen (from those displayed on the 'Time series' pane). Next, one of several available statistical methods (such as, for instance, moving averages, or a regression EM variant, etc.) must be chosen, and a time window including the values on which to base the estimate specified. Before actually starting computations, all settings are summarized for validation (Figure 13); after confirmation, an updated 'Time series' pane including the estimated value(s) – forecasts for years 2000 and 2001 for ISIC 3112 – is displayed as shown in Figure 14.

Figure 14. Updated Time Series Display

Clicking on the overlaid regression line (or estimated values) reveals further computational details of the derived estimate (Figure 15). At this point, one might decide to save the predicted values in the database for later recall. At present, the client admits to save data values only one by one; hence, after selecting an estimated value, another pane (Figure 16) opens on the screen. In this pane, all relevant documentation elements are listed, including the annotations generated (note especially the lower part of the pane) automatically. Furthermore, in this place an optional footnote can be entered manually and saved together with the data value as well as all other annotative

information generated. After that, the newly inserted data value can be retrieved from the database as usual.

Figure 15. Method and Computation Summary of Time Series Prediction

Figure 16. Data Imputation Save Dialogue Pane

This quick tour could of course not address all implemented features of the program prototype that, no doubt, leaves much to be desired from a practical usage point of view. Conversely, however, software design has taken a lot of care to not exclude various

enhancements and additions (such as more sophisticated prediction methods using multiple time series as input, etc.). At any rate, further development of the prototype certainly depends on practical experiences and requirements.

5 Concluding Remarks

As a key contributor of international statistics in the domain of industrial statistics, UNIDO takes the position of an information intermediary collecting ready-made (national) data either directly from national institutes or precompiled from the OECD. A major part of UNIDO's business concerns a variety of activities towards enhancing data consistency and especially long-term and cross-country comparability. To this end, reported data are screened and, if necessary and achievable, adapted either by consulting secondary data sources or by substituting doubtful or missing data with estimates based on econometric models. Nevertheless, deviances from international definitions cannot always be resolved entirely or satisfactorily, giving rise to additional comments for the sake of better data interpretation or to prevent misguided data usage.

Tougher user demands, cut budgets, economic globalization, diversifying data dissemination media and channels (notably, Internet-based services), and – last but not least – faster service cycles have led up to a situation where traditional business procedures can hardly be sustained any longer. As a consequence, also UNIDO moves into a re-design of its information infrastructures, drawing on recent developments and proposals in the statistical meta-information management domain. As one tile of the puzzle, the INDSTAT Databases of UNIDO have been singled out as a test-bed for the integration of statistical time series data and accompanying documentation. At the heart of this endeavor, the emanating information framework ought to achieve two primary goals, viz.

- help establish and maintain a minute documentation of all the updating and statistical procedures that incoming data undergoes at UNIDO, and
- assure a complete as possible documentation as to all shortcomings still affecting the data despite all efforts invested in removing them.

In response to these ambitions, based on a straightforward theoretical foundation, the model of a statistical information system has been presented that, regardless of its generality, seeks to address the specific difficulties and intricacies of a data body like the INDSTAT Databases. Consequently, system design stresses both, a structural data management component in support of data update operations interlinked with update documentation by recording the statistical methods applied and keeping track of the chronological order of operations, and a customized annotation management safeguarding the intimate linkage of data and related comments throughout all subsequent retrieval operations.

At the time of writing this report, a fairly worked out INDSTAT re-design has been accomplished, and a prototype implementation of part of its functionality successfully demonstrated. Some of the special problems of INDSTAT data (such as the irregularities stemming from ISIC code collapsing by reporting countries), while solved in principle in this research, have not been discussed here; other aspects – in particular a

general query engine coping efficiently with multi-hierarchical data (that is, data and annotation cubes exhibiting classification hierarchies simultaneously in several dimensions) – still await further theoretical and practical scrutiny. Finally, it also remains to be seen whether the proposed approach towards integrating aggregate data with its documentation beneficially extends to similar application domains.

References

- M.C. Chen, L. McNamee, and M. Melkanoff. A model of summary data and its application in statistical databases. In M. Rafanelli et al., editors, *Proc. Statistical and Scientific Database Management (4th SSDBM)*, pages 356–372, Berlin, 1989. Springer (LNCS 339).
- K.A. Froeschl und T. Yamada. The UNIDO Industrial Statistics Information Exchange Architecture (An Integrated Statistical Data and Data Documentation Framework). Working Paper No. 19, Work Session on Statistical Metadata (METIS), Conference of European Statisticians (Washington D.C., 28–30 November 2000), UN/ECE, 2000.
- R. Kimball. *The Data Warehouse Toolkit*. Wiley, New York, 1996.
- M. Rafanelli und F.L. Ricci. Mefisto: A functional model for statistical entities. *IEEE Transactions on Knowledge and Data Engineering*, 5(4):670–681, 1993.
- M. Silver. The role of footnotes in statistical metainformation systems. *Stat. J. UN/ECE*, 10(2):153–170, 1993.
- UN. *International Recommendations for Industrial Statistics*, United Nations, Statistical Papers, Series M, No. 48, Rev.1, Par. 7, 1983.
- UNIDO. *Industrial Statistics 4-digit Level of ISIC-code*. CD-ROM edition, UNIDO Statistics and Information Networks Branch, 2001.

Authors' addresses:

Ao.Univ.Prof. Dr. Karl A. Fröschl
Department of Statistics und Decision Support Systems
University of Vienna
Universitätsstraße 5
A-1010 Vienna

Tel. +43 1 4277 / 38614
Fax +43 1 4277 / 9386
E-Mail: karl.anton.froeschl@univie.ac.at

Dr. Tetsuo Yamada
Statistics and Information Networks Branch
Investment Promotion and Institutional-Capacity Building Division
United Nations Industrial Development Organization (UNIDO)
Vienna International Centre
P.O. Box 300
A-1400 Vienna, Austria

Tel. +43 1 26026 / 3667
E-Mail: T.Yamada@unido.org

Roman Kudrna
Hasenleitengasse 5/14/4
A-1110 Vienna, Austria