STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working Paper No. 19
English only

CONFERENCE OF EUROPEAN STATISTICIANS

UN/ECE Work Session on Statistical Metadata
(Washington, D.C., United States, 28-30 November 2000)

Topic (iii): Needs and responsibilities of international organisations for metadata

THE UNIDO INDUSTRIAL STATISTICS INFORMATION EXCHANGE ARCHITECTURE: AN INTEGRATED STATISTICAL DATA AND DATA DOCUMENTATION FRAMEWORK

Submitted by University of Vienna and UNIDO[1]

Invited paper

I. INTRODUCTION

1. In the context of globalization, international statistics for many social, economic and environmental areas that are, by their nature, collected, maintained and disseminated by international agencies have been increasingly demanded not only for cross-country analysis but also for country-specific analysis.

2. The usefulness of a multi-country database depends on the extent of cross-country comparability of its stored data. Thus, to assure sound use of their cross-country statistical data, international data sources are responsible for providing users with, in addition to general information for the overall database, relevant and detailed qualitative information that indicates the applicability and limitation of their disseminated individual statistical data in terms of international comparability. This is particularly important for such rather detailed sectoral statistics as the statistical outcome of national periodical surveys of manufacturing establishments, since they tend to be influenced largely by employed survey scope and statistical methodology, concepts, definitions and classification. Therefore, users need to know country-specific information with regard to those aspects from the viewpoint of international comparability (e.g., information concerning the deviations from international standards or norms).

3. In collaboration with the Organisation for Economic Co-operation and Development (OECD), the United Nations Industrial Development Organization (UNIDO) assumes the sole responsibility for the collection, development, maintenance and dissemination of worldwide key industrial statistics. As in the cases of other statistical data providers (national or international), the organization increasingly faces a necessity to adapt its statistical data management procedures in response to rapid changes in surrounding informational environment. This was recognized particularly through the organization's recent practices of licensing third parties for secondary dissemination of the UNIDO statistical databases and of web-site presentation of its industrial statistics and indicators.

4. Previously, owing to the traditional data dissemination procedures, it was quite possible to keep direct and regular contacts not only with national primary data suppliers but also with end users of disseminated data to communicate important background information for the conveyed data. In recent years, however, due to the IT revolution, more and more statistical information are transmitted to end users without opportunities of such direct interaction between the data disseminator and the end users.

---

[1] Prepared by Karl A. Froeschl, University of Vienna, and Tetsuo Yamada, United Nations Industrial Development Organization (UNIDO).

5. The crucial drawback of the traditional way of managing statistical background information is its apparent disintegration traditionally bridged, to a great deal, by "grown" staff expertise. This needs to be replaced with a system which is more consistent, flexible and sustainable.
The extent of complexity, and thus that of difficulty for its development, of a consistent, flexible and sustainable system of statistical background information depends to a large extent on the structure of the concerned statistical database system.

II. UNIDO INDUSTRIAL STATISTICS DATABASE SYSTEM

6. The Statistics and Information Networks Branch of UNIDO maintains a long-term data source called INDSTAT. Basically, the database system is a collection of annual time series on 14 selected economic statistics pertaining to manufacturing and covering 175 countries, with many of the series referring to past several decades, together with comprehensive computer software for data collection, maintenance and distribution, both in mainframe and PC computing environments.

7. Currently, INDSTAT comprises three industrial statistics databases*: Industrial Statistics at the 3-digit Level of ISIC(Rev.2), Industrial Statistics at the 4-digit Level of ISIC(Rev.2)* and *Industrial Statistics at the 3- and 4-digit Levels of ISIC(Rev.3).* Each of these databases is organized in accordance with three, four or five "stages" of data compilation. The first layer contains the data that were officially reported by national statistical offices (NSOs); the last layer cumulates the data contained in the previous layers and estimates made by UNIDO. The intermediate layers add data of decreasingly authoritative sources. The layout allows to retrieve the data according to the degree of confidence they deserve. With regard to statistical background information for INDSTAT, information for only data source, data class (or "stage") and for ISIC category combination[1] is embedded in the three databases.

8. Preparation of appropriate statistical background information (metadata) in support of INDSTAT databases requires relevant and concrete metadata inputs from primary data compilers (e.g., NSO). UNIDO requests NSOs to provide, together with statistical data, such inputs (country notes, table notes, footnotes) with regard to reported national data through its industrial statistics country questionnaire. The provided metadata are sometimes not described from the viewpoint of international comparability but rather from the viewpoint of national standards. In such cases, the organization re-describes/re-arranges the provided metadata into explicit information for the deviation from the international standard. This is often a difficult task and requires additional meta-information from the concerned NSO. Naturally, national statistical practice is based on the country's specific needs in legal, economic and policy aspects. At the same time, countries have been increasingly introducing international statistical standards because of the reason stated earlier, which result in less metadata requirements on international statistics and in more relevant metadata inputs from NSOs in terms of international standards.

9. Prepared metadata are then organized into a PC document (e.g., country notes) or a mainframe-based footnote database according to the nature of the metadata. At the same time, UNIDO documents the data development methods employed by the organization, data coverage and database structure of INDSTAT for dissemination purpose. Currently, however, both retrieval and update of those metadata are in many cases made rather clumsily on a manual basis.

10. Unlike the case of statistical data, consistent arrangement of metadata for dissemination/publication of any multi-dimensional (e.g., country, year, variable, industrial category, data class) statistical database is cumbersome if it is to be done manually: Different format and description are required for different purposes (e.g., different selection with regard to reference years, levels of industry classification, industries, data items, etc. among different INDSTAT-data dissemination media[2]).

---

[1] ISIC category combinations are, for some cases, not avoidable due to a mismatch between the national classification and ISIC or the statistical confidentiality role.

[2] 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

11. The rationalization of statistical metadata management calls for the development of a consistent, flexible and sustainable metadata system. Taking the above consideration into account, the Statistics and Information Networks Branch of UNIDO initiated in 1999 a project, as the first step, for the conceptual development and consequent designing of a statistical metadata system that would be suitable to INDSTAT technically and economically.

III. META-INFORMATION SYSTEM DESIGN PRINCIPLES

12. Having in mind the inherent structural complexity of the data bodies involved, only a comprehensive metadata-based system re-design approach has been considered promising at all. Thus, the project favors an *integrated data and data documentation framework* emphasizing that, while allowing scrutiny of data documentation (statistical metadata) both individually and jointly with statistical data, any statistical data access always entails the retrieval of associated metadata without demanding specific inquiry measures or actions. This way a rather tight interrelation of data and data documentation is both enforced and assured by purely technical means. However, as its major precondition, this principle presupposes a homogenous representation of all pieces of data documentation in order to enable uniform data and documentation access procedures. Moreover, as a change in data representation must not disrupt established UNIDO data services, a smooth migration policy is called for, leaving interface requirements of downstream systems and data usage almost untouched. Implying such a great effort to UNIDO, an expected side-benefit of re-designing the INDSTAT system is its potential applicability to further operational data management areas in need of refashion.

IV. FORMAL DESIGN FEATURES OF THE INTEGRATED FRAMEWORK

13. Basically, overall framework design centers around the notion of a "data cube" (now well-known from the field of data warehousing; e.g. cf. Kimball, 1996). Simply speaking, such a data cube resembles a multi-dimensional (cross-sectional) statistical table with cells each holding the value of some indicator (aggregate value, or macro-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 break-down as well as dimensions for separating different data classes (processing stages) and even different types of indicators. Clearly, each such dimension entertains its particular semantics and must thus be treated differently from the processing point of view. In a sense, this "enlarged" data cube might be conceived simply as a peculiar kind of spreadsheet where to fill in all stored data. Secondly, cube cell content distinguishes between (macro-)data and annotation data, letting 'annotation' denote any kind of remark or piece of documentation associated with a data cube cell. To allow for a subject-matter subdivision of annotations, the formal cross-classification concept even extends to specific "annotative dimensions" distinguishing different annotation classes (cf. Silver, 1993).

14. Formally speaking, the proposed information system architecture comprises two cubes, one for statistical data (the *data cube* proper) and another for annotations (the *annotation cube*), interrelated by a set of shared table dimensions. More specifically, let $d_1,\ldots,d_p$ denote the set of $p > 0$ formal table dimensions of the data cube; let further $B_i \equiv B(d_i)$ denote the basic value domains (domains of atomic values) associated with dimension d_i $(1 \le i \le p)$. On top of these basic domains, grouping functions $g_i : C_i \to 2^{B_i}$ are declared, mapping values introduced depending on the semantics of dimension i to non-empty subsets of B_i such that, in all cases, $B_i \subseteq C_i$, that is, all basic values of B_i re-appear in C_i.

---

tables (website), sales versions of INDSTAT (CD-ROM), pre-filled *General Industrial Statistics Questionnaire* (hard copy and Excel format). Currently, data users' access to available country-specific metadata is limited only to the Yearbook.

Now, each p-tuple $\mathbf{x} \in \bigtimes_{i=1}^{p} C_i$ identifies a (possible) macro-datum of the data cube, the value of which is referred to as $d(\mathbf{x})$; in database terms, $\mathbf{x}$ acts as an access key. Apparently, the value type of $d(\mathbf{x})$ depends particularly on the statistical indicator referenced by "category" (Chen et al., 1989) $\mathbf{x}$.

15. Depending on dimension and indicator semantics, various integrity constraints might be imposed on related $\mathbf{x}$-values. Naturally, the data space (cf. Rafanelli and Ricci, 1993) $\mathbf{D} \equiv \bigtimes_{i=1}^{p} C_i$ 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 $\mathbf{x}' \prec \mathbf{x}'' \Leftrightarrow g_i(x_i') \subseteq g_i(x_i'')$, $1 \le i \le p$ and $g_j(x_j') \subset g_j(x_j'')$ for some $j \in \{1, \ldots, p\}$. Now, for reasons of consistency, in case of a summarizable indicator, for $\mathbf{x}' \prec \mathbf{x}''$ it must necessarily hold that $d(\mathbf{x}') \le d(\mathbf{x}'')$. For strictly additive indicators, storage redundancy can be avoided by entering only those categories into the data cube not obtainable from stored values by summation.

16. With respect to the annotation cube, the set-up is entirely analogous except that, normally, at least one further "annotative" dimension is added. Thus, an annotation cube comprises the dimensions $d_1, \ldots, d_p, d_{p+1}, \ldots, d_q$ for $q \ge 0$, giving rise to its annotation space $\mathbf{A} \equiv \bigtimes_{i=1}^{q} C_i$ with B_j, C_j, and g_j ($p+1 \le j \le q$) defined accordingly. Likewise, each q-tuple $\mathbf{y} \in \bigtimes_{i=1}^{q} C_i$ identifies an individual cell of the annotation cube, the value of which is referred to as $a(\mathbf{y})$. Note that $a(\mathbf{y})$ pools *all* types of documentary stuff relating to $\mathbf{y}$, implying that a further break-down of material calls for the introduction of additional annotation classes (that is, an expansion of some B_j, $p+1 \le j \le q$) or even of a whole new annotative dimension itself. Quite in contrast to the data cube, the implication $a(\mathbf{y}') \Rightarrow a(\mathbf{y}'')$ 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.

17. By design, data and annotation cubes are interrelated easily through defining a projection of elements $\mathbf{y} \in \mathbf{A}$ onto $\mathbf{D}$, viz. $p(\mathbf{y}) \equiv p(y_1, \ldots, y_q) = (y_1, \ldots, y_p) \in \mathbf{D}$. Correspondingly, some annotation $y \in \mathbf{A}$ is "of relevance" for a datum $\mathbf{x} \in \mathbf{D}$ if and only if $p(\mathbf{x})$ and $\mathbf{y}$ intersect, that is $g_i(p(\mathbf{y})_i) \cap g_i(x_i) \ne \varnothing$ for $1 \le i \le p$.

18. Given this "twin cube" system design, query processing becomes quite a straight matter. A pure documentation query $\Omega_{\mathbf{A}} = \{w_1, \ldots, w_r\}$, $r \ge 1$, amounts to check whether there are annotations $a(\mathbf{y}_1), \ldots, a(\mathbf{y}_r)$ stored (that is, non-empty cube cells) for which either $w_k = \mathbf{y}_k$ or $w_k \prec \mathbf{y}_k$, $\mathbf{y}_k \in \mathbf{A}$ ($1 \le k \le r$). Owing to summarizability conditions, data queries might turn slightly more complex. For instance, assuming strict additivity, for any disjoint union of $\mathbf{x} \in \mathbf{D}$, say $\mathbf{x} = \mathbf{x}' \oplus \mathbf{x}''$, $d(\mathbf{x}' \oplus \mathbf{x}'') = d(\mathbf{x}') + d(\mathbf{x}'')$ subject to simple conditions such as $x_i' = x_i''$, for some i or the other, etc. Hence, generating a response to a (data) category $w \in \Omega_{\mathbf{D}}$ typically amounts to find a disjoint subdivision $\{\mathbf{x}_1, \ldots, \mathbf{x}_u\} \subseteq \mathbf{D}$ such that (i) all $d(\mathbf{x}_k)$, $1 \le k \le u$, exist in the database, and (ii) $\mathbf{x}_1 \oplus \cdots \oplus \mathbf{x}_u = v \in \mathbf{D}$ with v as "close" as possible to the queried w (it might be decided that $v \prec w$ all the time). For $u > 1$, the response value simply evaluates to $\sum_{k=1}^{u} d(\mathbf{x}_k)$; however, since determining v 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 data category $w \in \Omega_{\mathbf{D}}$ gives rise to a set-valued documentation query $\Omega_{\mathbf{A}} = \bigcup_{k=1}^{u} H(\mathbf{x}_k)$ where $H(\mathbf{x}) = \{\mathbf{y} \in \mathbf{A} \mid p(\mathbf{y}) = \mathbf{x}\}$. Contrary to pure documentation queries, in data queries no further

constraints must be imposed on the induced $\Omega_{\mathbf{A}}$ with respect to dimensions $p+1,\ldots,q$ (since this would cut off pertinent documentation elements). Apparently, in case that $v \neq w$ there is no point in retrieving annotations to parts of w actually not covered by the generated response v (thus, $\Omega_{\mathbf{D}} = \{w\}$ and $\Omega_{\mathbf{A}} = \{H(w)\}$ could in fact produce different documentary responses!).

IV. PUTTING THE FRAMEWORK INTO THE INDSTAT CONTEXT

19. The formal framework outlined assumes the following definite shape in the context of INDSTAT. The data cube is composed out of five dimensions, viz.

- a temporal break-down (in years);
- a geographic break-down (countries);
- a break-down of data in terms of ISIC (both Rev. 2 and 3);
- a formal break-down of data according to data class (processing stage);
- another formal break-down of data distinguishing between the (14) different economic statistics maintained within INDSTAT.

20. Of these five dimensions, probably most interesting is the set-up of the grouping function for the ISIC hierarchy of values. As with any classification tree, first the set B_{ISIC} of basic values is identified with all lowest-level classes 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). Because of the linear ordering of processing stages, the data class dimension actually does not get a grouping function attached (that is, $C_{\text{stage}} = B_{\text{stage}}$); typically, data values are passed on from one stage to the next. Likewise, customarily, the indicator dimension is not equipped with any groupings although, in the longer run, the expressive potential to state semantic relations between different economic statistics will certainly get used. Note, by the way, that obtaining a "computed" statistic from stored ones amounts, in this framework, to fill in derived figures into the data cube cells belonging to yet another processing stage (of same or different indicator)—just like filling spreadsheet cells by applying formulae expressed in terms of other cells.

21. To these data cube dimensions, the INDSTAT annotation cube adds one further dimension classifying all annotations in a simple scheme of some 10 distinct note classes adapted, basically, from an existing subdivision of country notes. There are note classes such as 'table note' ("ordinary" footnotes), 'source of information' (basically, background information about the NSOs providing ISIC data), 'definition remarks' (used for commenting on indicator etc. definitions not fully in line with established international standards), and so on. Annotation representation becomes in fact fairly parsimonious because of the powerful subsumption principle. For example, a peculiar concept definition used in a single country (or year, etc.) needs to be stated once only in the whole (annotation) database since, by subsumption, any single datum relating to this country (or year, etc.) "inherits" this note implicitly and gets thus retrieved in connection with any data value pertaining to this country (or year, etc.).

V. PRACTICAL IMPLEMENTATION ASPECTS AND DEVELOPMENTAL STATE OF PROJECT

22. In advance of populating the INDSTAT meta-information framework with real data and notes, its (slim) data model must be configured which in fact generates some preparatory overhead. However, this overhead pays off quickly since this way not only the notes attached to statistical data are properly maintained in the system: a great deal of structural information about INDSTAT data is represented formally as its *active-metadata* backbone for all subsequent query processing. In other words, the framework provides a neat and concise self-documentation directly amenable to formal application by information machinery.

23. At the time of writing this report, the UNIDO Statistics and Information Networks Branch has accomplished a fully worked out design of the INDSTAT meta-information framework. However, as yet neither the indicated retrieval routines handling data and documentation queries, nor any data migration schemes have been implemented.

REFERENCES

Chen M.C., McNamee L., Melkanoff M. (1989) A Model of Summary Data and its Application in Statistical Databases. In: Proc. Statistical and Scientific Database Management (4th SSDBM; Rafanelli M. et al., eds), Berlin et al.: Springer (LNCS 339), 356–372.

Kimball R. (1996) *The Data Warehouse Toolkit*. New York et al.: Wiley & Sons.

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

Silver M. (1993) The Role of Footnotes in Statistical Metainformation Systems. *Stat. J. UN/ECE* **10** (2), 153–170.