Role of Metadata in Quality Assurance of Multi-Country Statistical Data

In the case of UNIDO Industrial Statistics

Tetsuo Yamada, UNIDO

1. Introduction

In response to the rapid globalization in every socio-economic domain, international statistics for demographic, social, economic and environmental areas that are collected and disseminated by international data sources have been increasingly demanded not only for cross-country empirical analysis but also for country-specific ones. At the same time, the use of Internet for dissemination of those statistical data has resulted in increased user expectations concerning data availability, comparability and timeliness. While data dissemination through such medium has obvious advantages to both the data producer and users, it has resulted in diminishing opportunities of direct and regular interaction between the data producer and the users with regard to the quality and applicability of the provided statistical data.

Under these circumstances, data producers are expected more than ever to provide users with relevant information for the quality of the statistical data they produce and disseminate in a highly systematic and efficient manner following established standards such as the fundamental principles of official statistics and the Standards on Data and Metadata Exchange (SDMX). In the case of international organizations that maintain and disseminate multi-country statistical databases, relevant metadata for the assessment of the quality of the data contained in the databases are required by data users and at the same time by the organizations for their work on quality assurance of the data.

Probably the two most important quality dimensions of any multi-country statistical data are cross-country comparability and completeness of the data. Thus, international organizations’ quality assurance work on their respective databases focuses mainly on the assessment and improvement of data in terms of comparability and on data estimation to reduce data gaps.

However, even if a set of data reported to an international organization by a national data source is assessed as not comparable in terms of international standards, it is not easy in most cases for the international organization to increase the cross-country comparability of the data simply because the data were produced in accordance with national standards that deviated from international ones. Therefore, international organizations’ efforts in this area are often limited to only the assessment and documentation of the existing deviations of the data from international standards.

Taking the above points into consideration this paper focuses only on four quality dimensions of statistical data - international comparability, accuracy, consistency and timeliness - based on the UNIDO’s experience in the metadata system development for the quality assurance of its Industrial Statistics (INDSTAT) Database system. This is because the technical possibilities of
data improvement by UNIDO as a secondary data producer, as in the case of many other international organizations, are, to a large extent, limited to only two areas: (i) data estimation to increase the timeliness and completeness and (ii) screening and adjustment of incoherent or incomparable data reported by national statistical offices (NSOs). With regard to data accuracy, the organization’s quality assurance work relates only to data documentation.

In this context, this paper also intends to summarize (1) major sources of cross-country incomparability of existing data on industrial statistics and UNIDO’s efforts in ensuring the quality of those data, (2) roles of metadata in quality assurance of multi-country statistical data, and (3) organization of metadata in support of the INDSTAT Database and related metadata system design.

2. UNIDO INDSTAT Database System

There is an increasing consensus among development economists with regard to the argument that macroeconomic growth is the single most important necessary condition for poverty reduction. At the same time, in the context of globalization of economy, the notions of growth sources, comparative advantages and related structural changes have been increasingly significant in any macro-economic growth analysis. Thus, specialized and detailed economic statistics on industry (as well as on other economic sectors) are demanded more than ever by researchers and analysts to assess implications of the process of the globalization to individual countries. Furthermore, because of the nature of such growth analysis, the required statistical data must be internationally comparable. Any strategy, policy and programme for sustainable industrial development and investment that is not based on relevant empirical evidence would be subject to dispute. However, this argument holds only if the statistical data that support the empirical evidence are accurate, consistent, comparable and, needless to say, relevant and complete. Hence, an international data set with such quality standards on key industrial statistics should be available to all researchers, policy makers, business people and the like who are dealing with industry.

To support empirical analysis for industry and following the recommendation made by the UN Statistical Commission, UNIDO compiles, in collaboration with OECD, annual time series on key economic industrial statistics at detailed levels of industry aggregation (i.e., at 3- and 4-digit levels of ISIC). The compiled data are stored in the UNIDO INDSTAT Database system, supplemented with UNIDO estimates and disseminated worldwide. Throughout the process of the data compilation, those data are checked for consistency and coherency by both automated (or, computerized) and manual screening procedures. However, the assessment of the extent of the international comparability of those data in terms of statistical methods, concepts, definitions and classifications depends entirely on the related metadata provided by NSOs – primary data producers.

3. Potential Sources of Cross-Country Incomparability of Data and Ensuring Data Consistency

Perhaps the most important quality dimensions of a multi-country database are completeness (including timeliness) and cross-country comparability (including accuracy) of its stored data.
Data reported by NSOs are sometimes incomplete and not comparable across countries because they have been produced in accordance with the national statistical practice that is based on the country-specific needs in legal, economic and policy areas\(^1\).

In order to fill existing data gaps, the responsible international organization makes every effort to estimate missing data. On the other hand, for the improvement of cross-country comparability (and accuracy) of reported national data, what the organization can do is limited due to general lack of relevant supplementary statistical information. Modification of data reported by NSOs is further constrained because they are “official”.

Major determinants for the extent of international comparability of reported data on industrial statistics include scope of national establishment surveys, employed concepts and definitions of variables, and the national classification schemes\(^2\). For data accuracy, completeness and currentness of survey registers, good sampling and estimation methods, high response rates among surveyed units and appropriate adjustment of non-response are the basic requirements.

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 country (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\(^3\).

UNIDO statistical staff take considerable care in ensuring data consistency in the process of enlarging the INDSTAT databases, in improving the international comparability of their contents and in filling existing data gaps. 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 database system, to allow for possible corrections. The final screening takes place in two phases. First, possible abnormalities are identified through a computerized procedure. Second, the identified abnormalities are redressed to the extent possible.

---

1. An increasing number of countries have adopted recommended international standards in data compilation and consequently cross-country comparable data have been increasing gradually in recent years.

2. In the case of industrial statistics as well as many other economic statistics, they are measured/valued according to varying concept and definition. Thus, their information value is high but they are often incomparable to some extent across countries and sometime even over time.

3. 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 is used by the user to assess the quality of the data. Some countries fail to address the question altogether. The International Recommendations specifically request such information, and perhaps this is an area where improvements in national reporting practices may be anticipated.
In sum, UNIDO treats the national data reported by each NSO 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 by the NSO and returned to UNIDO, (i) manual and computerized 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) Filling data gaps: Imputation of missing data by utilizing data for related variable(s) or other available supplementary data; For selected variables, computer-based estimation of missing data by estimating inter-variable relations as country-specific time-series regression equations.

Achievement of the complete data coverage and, in particular, that of the increased cross-country comparability of data by employing common standards across countries are the top priorities in the maintenance of the INDSTAT database system. UNIDO promotes the international standard data-compilation methods and statistical concepts, definitions and classifications in the field of industrial statistics through its annual data compilation programme for industrial statistics and technical cooperation activities in the field of those statistics.

4. Role of Metadata in Quality Assurance of Multi-Country Statistical Data

Adjustment of incomparable data in line with international standards is often difficult as mentioned above. To ensure sound use of cross-country statistical data, the responsible international organization is expected to provide users with, in addition to general information for its database and its accessibility, relevant and detailed metadata that indicates the applicability and limitation of the statistical data in terms of accuracy and international comparability ranging from the country- and variable- to the cell-level. This is particularly important for detailed business statistics such as industrial statistics under the current consideration, since their data quality (particularly in terms of cross-country comparability) tend to be influenced largely by employed survey scope, methodology, concepts, definitions and classification schemes also as mentioned earlier.

Preparation of appropriate metadata in support of INDSTAT databases requires concrete and well-documented metadata inputs from primary data compilers (i.e., NSOs). 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 obtains meta information from NSOs 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 in the forms of, for instance, country notes, table notes, and footnotes provided by NSOs are often not described from the viewpoint of international comparability but rather from the viewpoint of national standards. In such cases, the organization is naturally expected by data users to re-describe/re-arrange the provided metadata into explicit information for the deviations of the reported data from the respective international standards. This is often a difficult task and requires supplementary meta-information from the concerned NSO.

The difficulties encountered during this exercise are often due to the descriptive nature of metadata: The possible sources of the difficulties include the facts:

(i) that responding officer of the NSO does not know the national standards in detail on which statistical data are based;
(ii) that he/she does not know in detail the corresponding international standards; and
(iii) that, due to lack of common statistical terminology in the international language(s) as well as in the working language of the NSO with one-to-one correspondence between the national and international languages, it is difficult for both the NSO and the international organization to communicate with each other properly for metadata exchange.

To overcome at least partially those difficulties, UNIDO has introduced, as part of the UNIDO country questionnaire, a metadata questionnaire to be completed by NSOs which is designed to enable the organization to obtain explicit information as much as possible for existing deviations of the statistical data reported by NSOs from international standards. To complete the questionnaire, NSOs are requested to tick relevant items presented in the questionnaire in accordance with their national standards. For instance, in the case of the data item, output, in the context of industrial establishment surveys, there are basically two concepts (census and national account concepts) and two valuations (at factor costs; at producers' prices) and their several variations, and hence, there are minimum four different definitions. In order to cover these four definitions and possible deviations from them, the questionnaire asks the NSOs to tick all relevant components of the nationally defined output (internationally recommended output is “census” output at “producers’ prices”) as shown below (Items in Italic are the components excluded in the international standard definition of census output).
value of all products
net change between the beginning and end of the year in the value of work in process and stocks of goods to be shipped in the same condition as received
value of industrial work done or industrial services rendered to others
value of goods shipped in the same condition as received less the amount paid for these goods
value of fixed assets produced by the unit for its own use
net change between the beginning and end of the year in the value of stocks of finished goods
subsidies
indirect taxes
revenue from non-industrial activities
others (Please specify): ________________________________

Depending on the type of industry, different concepts or valuations make significant differences in actual data. For instance, cigarette manufacturers spend a huge amount on advertising their products. Due to the large expenditures relating to the non-industrial service, the difference between the output (as well as value added) measured in accordance with the census concept and that in the national account concept for the cigarette industry is relatively large. On the other hand, for some basic-need-related industries, government subsidies may be significant, which results in a significant difference between the output (and value added) measured at factor cost and that at producers’ prices.

Metadata that is arranged in this sort of form enable the international organization to re-describe it in a way that it would indicate explicitly the existing deviations from international standards and, therefore, reduce considerably the need of international organizations’ data queries to NSOs which in turn reduces the data-reporting burden on the NSOs. Furthermore, it would result in consistent presentation of metadata in a statistical website and other data-dissemination media.

5. Metadata System Design – the case of the UNIDO Industrial Statistics database system:

The UNIDO Industrial Statistics Database (INDSTAT) system consists of basically three databases: INDSTAT at the 3-digit level of ISIC-Rev.2 (INDSTAT3) which contains long time series; INDSTAT at the 4-digit level of ISIC-Rev.2 (INDSTAT4); and INDSTAT at the 3- and 4-digit levels of ISIC-Rev.3 (INDSTAT-REV3). In each of them, a data cell is specified by four keys: country, year, table (variable) and ISIC category.\(^5\)

\(^4\) Here again, the need of the diffusion of common terminology among national and international data compilers is apparent.
\(^5\) Actually, there is one more key called “stage of internal data processing”. However, for the sake of simplicity, this fifth key is ignored in this paper.
As in the cases of other statistical data providers (national and 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 practices of licensing third parties for secondary dissemination of the UNIDO statistical databases and of web-site presentation of its industrial statistics and indicators.

The crucial drawback of the traditional way of managing metadata in support of the statistical data is its apparent disintegration traditionally bridged, to a great extent, by staff expertise. ICT-based data dissemination (e.g., on-line databases, web-based and CD-ROM data-dissemination products) together with imbedded metadata-related interface requires a system that is more consistent, flexible and sustainable. Unlike the case of statistical data, consistent arrangement of metadata for dissemination/publication of any multi-dimensional 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, aggregation levels of industry classification, industries, data items, etc. among different data dissemination media). For instance, metadata for each of the three INDSTAT databases can be structured into a hierarchy in accordance with combinations of the four keys but rather in an arbitrary manner from the top (e.g., country-level notes) to the bottom (e.g., cell-level footnotes). Difficulties in organizing and maintain the metadata stem from the fact that they are subject to re-described and shifted in the hierarchy over time or depending on the selection of a subset of the statistical data in accordance with the four data dimensions if the associated metadata are to be presented consistently. Such a shift occurs when original description for the metadata becomes irrelevant and thus needs to be re-described.

The rationalization of statistical metadata management calls for the development of a consistent, flexible and sustainable metadata system. Taking the above consideration into account, 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 technically and economically to the UNIDO INDSTAT system.

The project for the development of an integrated data and data documentation framework emphasizes 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 (i.e., one-to-one correspondence at the cell level being specified by the four keys (or dimensions) – country, year, variable, classification category) 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.

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. 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 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).

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.

The formal framework outlined assumes the following definite shape in the context of INDSTAT. The data cube is composed of five dimensions as stated above:

- 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 break-down of data distinguishing between the different industrial statistics maintained within INDSTAT system.

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 industrial statistics 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.).

A practical version of the conceptual metadata system is currently employed in the CD-ROM dissemination product of INDSTAT-REV.3 and in the production of a hardcopy sales publication driven from the INDSTAT database system, the *International Yearbook of Industrial Statistics*.

---