



Committee for the Coordination of Statistical Activities

Conference on Data Quality for International Organizations
Newport, Wales, United Kingdom, 27 - 28 April 2006



Session 5: Tools and practices for collecting quality metadata from data providers and informing users on the quality of statistics

**ASSESSMENT OF STATISTICAL DATA QUALITY:
THE EXAMPLE OF THE OCCUPATIONAL WAGES PART OF THE ILO OCTOBER
INQUIRY**

Presentation of the October Inquiry¹

The ILO October Inquiry (OI) is a worldwide survey of wages and hours of work relating to 159 pre-selected occupations within 49 industry groups, and of retail prices of 93 food items, conducted with reference to the month of October of each year. It was initiated in 1924 to give effect to a resolution of the 1st International Conference of Labour Statisticians (1923), with the initial objective of making comparisons of real wages and of the food purchasing power of wages for workers in member countries. On the recommendation of various International Conferences of Labour Statisticians (ICLS), the occupational scope of the Inquiry and the types of wages and hours covered were progressively extended, and the number and types of articles for which prices were collected were progressively increased. The last major revision and expansion of the Inquiry was introduced in October 1985.²

The information is collected by means of two questionnaires, one relating to wages and hours of work and the other to retail food prices, which are addressed to national statistical offices (NSOs) and/or relevant ministries. The ILO does not expect reporting agencies to conduct special surveys in order to complete the questionnaires, but to supply whatever relevant information is already available. In the case of wages and hours of work, a number of sources are commonly used, including regular occupational wage surveys, labour force surveys, administrative records of wage-fixing bodies and regulations. Data are collected on wage rates, earnings, normal hours of work and hours actually worked or paid for³, by sex, where available. In general, the retail prices reported are the averages of

¹ *Occupational Wages and Hours of Work and Retail Food Prices – Statistics from the ILO October Inquiry*, available on line: <http://laborsta.ilo.org>.

² For an explanation of the origin, development and revision of the ILO October Inquiry, see ILO: *Bulletin of Labour Statistics – October Inquiry Results, 1983 and 1984* (1985) and "The Revision of the ILO October Inquiry: Retail Prices Part" in *Bulletin of Labour Statistics*, second quarter 1986.

³ For the definitions of these concepts, see the full text of the relevant Resolutions in *Current international recommendations on labour statistics* (ILO, Geneva, 2000) or the ILO Bureau of Statistics' web site at : www.ilo.org/stat/.

those collected for the purposes of calculating consumer price indices. Some countries, however, do carry out special data collections for some items which are not normally covered in the regular pricing programmes.

The survey results are disseminated without adjustment, except in the case of retail prices when some conversions are necessary in order to be able to publish prices for standard quantities. Despite efforts to promote the comparability and continuity of the data, there are some unavoidable differences between the concepts used, specifications of occupations and items, reference periods, types of sources and methods of data collection in the various countries.

Steps taken by the ILO to ensure data quality and comparability

Detailed descriptions of the occupations are provided to the reporting agencies, with a view to helping respondents identify the respective occupations, and enhancing the comparability of data overtime and between countries. Each description is given in the form of a short opening statement describing the general functions of the occupation, followed by an enumeration of the main tasks usually performed. The description may mention possible variations in the way in which the work is performed, and may include certain tasks which are sometimes performed by workers in the occupation but which are not necessarily considered as inherent in that occupation.

The occupations and industry groups have been coded, as far as possible, according to the International Standard Classification of Occupations (ISCO) and the International Standard Industrial Classification of All Economic Activities (ISIC) respectively. The codes for the 1968 and 1988/90 versions of each classification are listed at the end of the descriptions.

Detailed instructions are also provided to the respondents. In the case of wages and hours of work, the information should relate to adult full-time employees who are fully qualified, i.e. who have acquired the training and experience normally necessary for the occupation in question. The relevant definitions of wage rates, earnings, normal hours of work and average hours of work per week (hours actually worked or average hours paid for) are included in the covering pages.

Detailed descriptions of the food items are also supplied along with the OI questionnaire, which include suggested pricing units (i.e. the weight, volume or number of each item for which the price should be reported) and for four items (bread, cheese, oil and fish), countries are asked to specify the particular variety.

The questionnaires include an introductory information section for each part, which respondents are required to complete, indicating or confirming the geographical and employee coverage of the statistics, the reference period (if other than October), the currency and in the case of change of currency, the equivalence between the old and the new currencies; the source of data (name of survey, type of administrative record, etc.); the publication, if any, in which the data appear; the department or agency completing the questionnaire, as well as the details of the person who may be contacted for further information.

For the last four years, the questionnaires have been available in three formats: (i) an electronic version accessible from the ILO/LABORSTA web server, with country-specific passwords and relevant instructions; (ii) an Excel-type executable file (.exe); or (iii) a printed version. The electronic questionnaire is the method preferred by the ILO Bureau of Statistics: it allows for computerized uploading of the data into the database, thus avoiding the risk of errors in manual data entry and improving the timeliness of data collection and dissemination; however analysis of the replies (scrutiny of data) and manual entry of codes and notes (metadata) are still required. In the case of Excel-type and printed questionnaires, manual processing of the data and metadata (notes) are still the rule.

The questionnaires, whether in electronic format or in printed form, contain room for two years of data: the current year (e.g. October 2005 collected in 2006) and pre-filled data for the previous year, so that the reporting agencies can check the statistics previously provided, and if required, amend or revise them.

Quality problems encountered in the Occupational Wages part of the OI

In principle, the OI would be an ideal source of information for comparing wages received by various groups of workers within a given country and across countries, or for assessing their purchasing power over time and between countries, provided each country provided information on wages in a consistent way and based on ILO definitions and instructions. In practice, they report wages in various and sometimes inconsistent ways.

As already mentioned, countries send the ILO data obtained from a range of different sources: household/labour force surveys; establishment occupational wages surveys; collective agreements; legally determined minimum wages or wage scales; or other types of surveys or inquiries of varying quality. Within a given country, data sources may change over time, thus providing different types of wage statistics, data for different occupations or statistics expressed in various time units. Some countries provide both wage rates and average earnings; others give wages for men and women separately, while others give wages for one sex and for both sexes, thus hampering comparability between countries.

Another difficulty lays in the fact that the wages statistics are collected for detailed occupations (at the three- or four-digit level of ISCO) within specific industries (at the two- or three digit level of ISIC). This type of cross-classification may be difficult to achieve, especially when the sample size of surveys (whether of households or establishments) does not provide for reliable and representative data at this fine level of classification. Thus, in some countries, the occupational wages are not classified by industry, but by industry group at a more aggregate level of the industrial classification.

Last but not least, the database has gaps in terms of both entire data series for some countries, and data items for some others. This is due to one of the following reasons: the data does not exist officially at the country level; it exists but could not be obtained; or the data was collected but a decision was made not to include it, based on the quality assessment of the reply.

Quality assessment undertaken by the ILO in 2001

Regular checks are made by the ILO Bureau of Statistics on the quality of the statistics received from reporting agencies: The OI does not escape the rules. The wages data are scrutinized and outliers or suspect downward or upward trends are identified. Comparisons are made with the general averages of wage rates, earnings and hours of work by branch of economic activity provided by these same countries for the ILO Yearbook of Labour Statistics, as well as comparisons of trends with the Consumer Price Indices. In all cases, efforts are made to request the reporting agencies to verify any dubious data and to provide complementary methodological information on replies that “look” questionable. In general, the “discrepancies” noted are due to methodological revisions, sample changes, different coverage of collective agreements, etc.. Whenever possible, differences between the required data and those provided by reporting agencies are indicated in footnotes attached to the series.

In 2001, in response to some major users of the OI database, both within and outside the ILO, the ILO Bureau of Statistics started an assessment of the quality of the data contained in the computerized database (1983-2000), allowing, through coding, differentiation between the data supplied. It classified the various sources and types of data into four quality groups, ranging from “not acceptable” (coded 1) to “excellent” (coded 4). This classification was based on the consistency of the data (regularity of compilation, trend of the wages data, consistency between wage rates and earnings, etc.),

and information available to the ILO on the source and scope of the data – independently from the use that could be made of the wages data.

The vast majority of ratings of data quality covered single countries, but in a few cases where the country used different sources or changed sources over time, multiple ratings were provided. The bulk of the data were rated as being in the “acceptable/good” category (52.6 per cent) or the “excellent” category” (32.4 per cent). These ratings (3 and 4) were generally applied to consistent data series mainly derived from regular household or establishment surveys carried out in statistically developed countries, as well as to series of wages data derived from collective agreements, labour legislation and similar administrative records.

Some 14 per cent of the data were considered as of “poor quality” and one source as “not acceptable”. The latter referred to a single year data set and was subsequently deleted from the database. “Poor quality” data were generally erratic data derived from various administrative sources or from ad-hoc data collections which were not considered as sufficiently representative and reliable at the specified level of detail: In some cases, the countries concerned had expressed difficulty in collecting the required data and in checking their validity, or indicated inconsistencies in the sample or regional coverage, the concepts and definitions used (e.g. wage rates instead of earnings), or else had not been in a position to check the computations of the wages (e.g. in the case of 10 or more year old statistics).

Consequently, a number of modifications were made to the OI database, and misclassified or miscoded data were corrected (e.g. according to concept or time unit). However, after verification with the countries concerned, the statistics considered as of “poor quality” were retained in the database, because these data had been published in the earlier versions of the October Inquiry, and also because for some analyses, data of dubious quality may be preferable to no data at all.

The October Inquiry statistics were subsequently further standardized to develop an “Occupational Wages around the World” (OWW) data file, using the above-mentioned quality rating⁴. The various statistics were calibrated into a normalized wage rate, using the data deemed of “acceptable/good” or “excellent” quality. Adjustment coefficients were calculated, that measure how non-standard forms of data diverge from the standard rate for different countries, occupations and time periods; and the wage rates of all occupations across countries were standardized based on the most common form of data in the Inquiry, the average monthly wage rates for male workers.

Similar and regular quality assessments were carried out by the ILO in the following years, with a view to strengthening the quality of the OI database, and providing support to users of these statistics – in particular for the update of the OWW dataset, and the compilation of occupational wage indices contained in the ILO Key Indicators of the Labour Market (KILM). These assessments essentially concerned the accuracy of the data, their timeliness and their coherence.

The next steps

The ILO is engaged in a revision and expansion of the October Inquiry. A preliminary examination of the importance of existing and potential occupations on the basis of trends and projections in labour force structures in various countries has been carried out, in consultation with ILO sectoral specialists and a number of outside users (international organizations, economists, etc.). This review should help increase the *relevance* of the statistics to meet various users’ requirements. Plans are also underway to collect the corresponding occupational employment data to assess the importance of the selected occupations.

⁴ For details on the use of the ratings and the standardization procedure, see: *The Occupational Wages around the World data file (OWW)*, Richard B. Freeman and Remco H. Oostendorp, in *International Labour Review*, Volume 140, Number 2001/4, ILO, Geneva; website: <http://www.ilo.org/public/english/support/publ/revue/index.htm>. The OWW data file is available from the NBER website at: <http://www.nber.org/oww>.

As regards the statistical *quality* of the data provided for the OI and released to users, it is planned to improve two mechanisms: (i) improved collection of metadata on this Inquiry, in particular with regard to administrative sources; and (ii) introduction of a number of computerized validation checks in the database, in order to systematically identify outliers, significant downward or upward trends, inconsistent time units, and similar quantitative measures which, at present, are still subject to visual scrutiny and manual edit.

One of the criticisms expressed about the OI refers to the lack of consistency of wages statistics per time unit. In an effort to standardize the OI database, reporting agencies are being asked to ensure that statistics of *hourly* wage/salary rates and/or earnings are provided whenever possible. If wage/salary rates or earnings are expressed in another time unit (e.g. weekly or monthly), they are requested to ensure that the corresponding number of weekly or monthly hours of work (normal hours, hours actually worked or hours paid for) are provided together with the wages data. This should allow the Bureau of Statistics to compile and disseminate if possible (i.e. with the reporting agencies' consent) more systematic hourly statistics, thus improving the cross-country *comparability* of the OI data.

* * * * *

ILO experience with gathering and disseminating meta-data on household income and expenditure statistics

Summary: This paper summarises ILO experience in the gathering, processing and dissemination of meta-data on household income and expenditure statistics.

Introduction

In late 2002 and early 2003, as part of its preparations for the 17th International Conference of Labour Statisticians, Geneva, 24 November to 3 December 2003, the ILO took action to update the information presented in the ILO's *Sources and Methods* publication that covers household income and expenditure statistics (HIES)⁵. The results from the last systematic updating of this publication were published trilingually in 1994.

Previous practice

The ILO continually monitors the methods used by countries to produce labour statistics. These methodological descriptions are included in the on-line statistical database LABORSTA (laborsta.ilo.org) as well as in the ILO's *Sources and Methods* publications. The methodological descriptions also form the basis for footnotes to national statistics presented in ILO's *Yearbook of Labour Statistics* and other statistical publications.

In addition to this, efforts are made at infrequent intervals to ensure that the methodological descriptions on a particular topic are complete and up-to-date for all countries. This updating is generally carried out in rotation for different topics.

In the past, the practice has been:

- (a) For countries that have previously submitted a description: to send them the previous text and ask for a revision or update;
- (b) For countries that have not previously submitted a description: to ask for a description of the source.

This process takes time and places a burden on both the ILO Bureau of Statistics and the national statistical offices.

The resulting descriptions are translated into the other two of the three working languages of the ILO (English, French and Spanish) and converted into electronic form for dissemination, both in print and in electronic form, e.g. on the ILO's websites. These descriptions do not lend themselves to easy analysis of differences in concepts, definitions and techniques.

Latest method

In an effort to reduce the burdens in reporting, processing, translation and reformatting for dissemination, the ILO Bureau of Statistics decided to use a new approach for updating the meta-information on national household income and expenditure statistics, in which:

- (a) A questionnaire was designed with tick boxes offering a choice of answers for all main features to the extent possible. (Offering a choice of answers simplified and speeded country reporting. As the design had to allow for all possible alternatives it required sound knowledge of likely methods. "Other, specify" options were provided throughout.);

⁵ In the division of labour between statistical units within the UN system ILO serves as the custodian for household income and expenditure statistics.

- (b) This questionnaire was tested in one country. This resulted in some revisions to the first draft;
- (c) The finalised questionnaire (amounting to over 80 questions plus annexes) was translated into French and Spanish;
- (d) By oversight, codes were not pre-printed alongside each tick box in the final version, and this slowed data entry, especially for those questions where there were many alternatives;
- (e) A set of complex computer programs were developed to:
 - i. Permit data entry into a database⁶ of codes representing country responses (and accompanying text where the tick box responses were not sufficient);
 - ii. Carry out automatic checks on data entry for valid codes, coherency between answers, etc to reduce errors during data entry;
 - iii. Allow extensive data analysis (cross-tabulations) of the country responses. These analyses were automatically updated as each new set of data was added to the database;
 - iv. Produce output in HTML format of the textual descriptions of each source in three languages. Again these were automatically updated as each new set of data was added to the database;
 - v. This HTML output could then be directly inserted into website descriptions or used in printed publications.
- (f) For each question, different output descriptions were prepared in the three languages reflecting the variations allowed for by the tick box format. In order to create a full description of a particular national HIES, these separate individual sentence elements were then concatenated in such a way that any source could be described. As indicated in (e)(iv) above, these elements were used to generate text in all three languages as part of the computerised output;
- (g) The database that was created allowed extensive cross-classification and analysis. (Significantly more than previous methods had permitted.) This database will be made available to external users on request.

Resources required

The resulting effort required:

- (a) one work month of questionnaire design and computer specification;
- (b) one work month of translation;
- (c) 3 work months of computer effort; and
- (d) initially one (later extended to two) work months of data entry.

Previous work by the Bureau of Statistics to generate output of this type required many months of clerical preparation of a methodological description for each country/source and then the translation of each of these into the other two languages. The translation process alone amounted to well over \$US 40,000 for each publication, and it was the extent of translation that the above computerised system was aiming to replace. In practice, the translation component of the textual descriptions under the new procedure amounted to some \$US 17,000, and the development of the computerized system, to some \$US 15,000, i.e. a total of \$US 32,000.

It should be noted that the investment in the computerized system will benefit the production of future similar outputs.

⁶ There were in fact six separate databases each covering a different part of the questionnaire, but for ease of presentation, this paper refers to one combined database. The six databases were linked by a common variable, which was the identification number of the methodological description.

Experience gained

In practice, it was found that most countries added comments and qualifications to what had been expected to be adequate response alternatives. This complicated and lengthened the data entry and added a considerable burden to the translation process.

As at September 2003, the data entry work was still in progress and the translation of these additional (unexpected) comments had not started.

It is believed, nevertheless, that the approach facilitated and improved country response and that the total translation burden was reduced with consequent savings for the ILO. In addition, the use of a computerised database permits a much richer analysis of methods and facilitates electronic dissemination to external researchers.

However, the approach required more care in designing the questionnaire and its pre-coded responses than with the previous approach. For the future, it would be prudent to design the questionnaire and database to allow comments in respect of all responses and to permit multiple responses to most questions. It has also been proposed that the resulting questionnaire might be prepared and dispatched to countries in an electronic format in order to avoid delays due to postage⁷.

Conclusion

Would we do it again? We are already doing so in preparing methodological descriptions for statistics on the employment situation of persons with disabilities, which however, are based on a much shorter, simpler questionnaire which has received similar types of replies. In balance, we believe that this approach has risks and high computing costs but it is worth repeating for the next round of methodological description updates, with details modified in light of the experiences gained.

⁷ The use of electronic questionnaires and dispatch may not be possible or effective in all cases. Occasionally, the contact e-mail address is not available and/or more than one agency is involved in completing the questionnaire.