UNITED NATIONS INDUSTRIAL DEVELOPMENT ORGANIZATION
Vienna International Centre, P.O. Box 300, 1400 Vienna, Austria
Tel: (+43-1) 26026-0 · www.unido.org · unido@unido.org

D03571

Distr.
LIMITED

ID/WG.106/8
3 September 1971

ENGLISH
Original: FRENCH

United Nations Industrial Development Organization

Training Workshop in Methods of Industrial Surveys
Dakar, Senegal, 13-24 September 1971

DATA PROCESSING [1]

by

Mr. Georges Godin
Director, INSEE
(National Institute of Statistics and Economic Studies)
Paris, France

[1] The views and opinions expressed in this paper are those of the author and do not necessarily reflect the views of the secretariat of UNIDO.

TABLE OF CONTENTS

I. Checking the availability and quality of data

The purpose of checking is three-fold:

(a) To establish that there is a completed questionnaire for every enterprise surveyed, that the type of questionnaire is compatible with the characteristics of the enterprise and that there is an answer to every question;

(b) To check that if two or more variables are linked by a relationship of equality or inequality, the replies meet these conditions;

(c) To correct omissions or anomalies encountered.

1.1 General comments on the method of checking

(a) No question can be left unanswered; if there is a blank, an "0" should be inserted or else an "on sight" assessment made, or a further survey should be carried out. Otherwise, provision should be made for automatic checking during electronic data processing.

(b) The consistency between the variables A, B, C, ... can be checked by the application of the following general formulas:

$$A = P; \quad A < P$$
$$A + B + C + . = P; \quad A + B + C + . < P$$
$$P = M \pm e \quad ; \quad A/B = R \pm e$$

The effectiveness of checking depends on the number of these correlations between the values. It is non-existent for a survey which has only unrelated variables: in such a case, checking can be done only by comparison with external data. It is significant in the case of an industrial survey.

(c) If, for a group of values, m_i, it is established that $M = \Sigma \ m_i$
If, for a group of values, n_i, it is established that $N = \Sigma \ n_i$
If, lastly, the condition $M = N$ is met, it can be deduced (for checking purposes) that each of the values has been calculated accurately.

Example: All the elements of operating and profit and loss accounts and balance sheets are considered correct if the operating surplus reappears in the profit and loss account and if the balance shown in this account is equal to that shown in the balance sheet. (There are also special conditions which must be met, for example, equal fluctuations of reserves according to both the balance sheet and the operating account.)

(d) If, for a group with the value p_i, it is established that $\Sigma p_i = P$. If the requirement $P = Q$ or $P < Q$ is met, Q being a value assumed to be correct, it can be concluded that each of the values p has been calculated accurately.

Example: An industrial survey provides for each establishment the total gross formation of fixed capital. As far as the enterprise is concerned, the gross formation of fixed capital may, thus, be calculated in two ways: either from the data for establishments or from the corporate fixed assets entered in the balance sheets. If, after a checking operation of the type described in paragraph (c) has been carried out, the elements of the balance sheets are assumed to be correct, and if the two methods of calculating the gross formation of fixed capital of the enterprise give the same results, it can be concluded that the gross formation of fixed capital for each establishment is correct.

(e) Let us suppose that for two values g_1 and g_2 it is established that $R - e < g_1/g_2 < R + e$. As a rule, nothing can be deduced from this unless, in addition, one of the two values is assumed to be correct; if that is so, however, the other value is also correct.

Example: The industrial survey provides, for each establishment, the total amount of salaries and social welfare payments. There are, thus, two methods of calculating the total of salaries and social welfare payments for the enterprise as a whole: either by using the data on establishments or from the relevant part of the operating account. The use of a checking system of the type described in paragraph (d) will lead to the acceptance or rejection of the total of salaries and payments for each establishment. If the total of salaries and social welfare payments of an establishment is assumed to be correct, the accuracy of each component can be tested by calculating the ratio $\frac{\text{social welfare payments}}{\text{salaries}}$ which should be included within a certain range of values based on social legislation and experience.

1.2 Preparation and execution of the control plan

Two practical questions should also be settled before the survey is launched:

- What form of checking should be done?
- Where when and how should checking be done?

(a) Among the variables which must be collected to produce the set of statistics required, one must identify pairs or groups of variables for comparison and define the conditions which must be met by the variables of the group. This operation will show whether it is necessary to add to the questionnaire a question which is of value only for checking purposes; such is the role of balances in operating and profit and loss accounts and balance sheets.

(b) The foregoing examples illustrate the need to establish the order in which checking operations are to be carried out. It will be noted that the first thing to be checked will be whether satisfactory questionnaires concerning the various units of the enterprise are available and whether there are replies to all the questions.

(c) All the checks will be carried out in the enterprise when the data is collected. They will be repeated when the data is processed by computer. If the first phase of the checking is conducted correctly, it will eliminate the need for additional inquiries for verification purposes (this causes additional expense and delay). The respondent should, therefore, be informed of all the checking operations which will be carried out before his answers are processed. Furthermore, all the interviewers must be supplied with the control plan and should ask respondents to explain the reasons for deviations from norms (the reasons to be noted in an annex to the questionnaire). The distribution of the control plan is the best way of keeping errors to a minimum. This is a decisive factor in reducing processing delays. It enables checking to be done at the enterprise (i.e. at the very source of information), and at the same time as the data is collected.

II. Manual operations and automatic processing

2.1 Preparation of questionnaires for punching

Two operations must be carried out at this stage:

(a) The preparation of the questionnaires and the making up of batches. This involves on the one hand making any simple changes required, such as making letters more legible, rounding off figures by removing decimals, if any, and on the other hand sorting the questionnaires into batches for numbering and punching. Generally speaking, a batch will consist of no more than 100 questionnaires. A general list will be drawn up, indicating the type and number of questionnaires to be included in each batch. The progress of work will also be noted on this sheet: start and finish of conversion into digits, start and finish of punching.

(b) Conversion into digits: To eliminate the need for supplementary documents for this operation, "number boxes" should be printed on the questionnaire. The operation consists of converting into numbers information which may be given in the questionnaire in numerical or alphabetical form.

Examples of data which will be coded numerically include the geographical area, the legal form, the activities and products, and the order of magnitude of the total number of staff and the turnover.

Obviously, nomenclatures and codes worked out when the survey was drawn up will come into their own at this stage.

Attempts are now being made to streamline this operation in two ways:

- By the use of precoded questionnaires. This means that, ideally, the data is coded at the time of the inquiry itself. The precoding of questionnaires consists of enumerating on the questionnaire all the items of a classification matched up with their respective code numbers. The respondent is asked (i) to select the relevant entry or entries and (ii) to give the appropriate reply; this applies to questions referring for example to legal form or to the description of products manufactured;

- By automatic codification, particularly when coding is done on the basis of numerical data. This applies to the codification of classes of size or the determination of the principal activity by analysing the distribution of staff or sales among various activities.

2.2 Punching

In spite of recent innovations (punched tape, direct recording on magnetic tape, optical reading), the punch card is still preferred as the main carrier of information to be processed by automatic procedures - traditional punch card equipment and now computers. The punch card most commonly used measures 8 x 18 cm and has eighty columns of data recorded as perforations in the columns, representing numerical or alphabetical characters or other symbols. Some key punches automatically print each character at the top of the card at the same time as the perforations are made, which facilitates reading.

Errors in punching are often more serious than those which occur in manual transcription (punching errors in the strict sense, inversion of columns, field shifting, etc.). It is essential that these errors should be detected if they occur. Two methods may be used:

- The first, which is traditionally used because it is the only method compatible with traditional punch card techniques, consists of repeating the operation on a checking machine which stops whenever there is a discrepancy between the record made by the operator and the record which appears on the card being checked;

- The second method of checking is by computer. It consists of ending the record with a control digit selected according to the content of the record. The computer checks whether the control digit is compatible with the record itself.

The design of the card or record clarifies the one-to-one correspondence between the headings on the questionnaire and the card file to be established. To ensure efficiency in punching, recording should preferably be carried out in the order of reading from the basic document. Card designs may differ according to whether processing is done by a tabulating machine or by computer, but attention should in any case be paid to the record length which may be fixed (in which case it can accommodate the largest numbers) or variable (in the case of chain punching, which can be done only when the data is to be processed by computer).

2.3 Processing by punch card techniques

This consists basically of sorting and tabulation operations.

The purpose of sorting is to arrange the card file according to the symbols representing the categories forming table divisions (geographical area, code of activity, product, size, etc.).

Tabulation by an automatic tabulator makes it possible to enumerate the total of the batches formed by sorting and to make totals at successive selection points. The results are printed on a paper tape. The distribution of characters, letters, numbers, signs and spaces must be studied in advance.

An operation generally includes several sorting and tabulation sequences.

2.4 Computer processing

A computer basically consists of two types of components:

- A central processing unit: memory, processing unit and control console;
- Peripheral control units: card readers, tape unwinders, printers.

The main factor determining the data processing performance of a computer is the capacity of its main memory. Its performance is measured in terms of the speed at which it can carry out sorting and summarizing operations, of its capacity for making at high speeds,calculations which may involve a large number of variables - conventional statistical calculations, inversion of inter-industry matrices, etc. - and of its capacity to store data for later use.

The processing of data by computer imposes, however, rigorous constraints which cannot be ignored without certain failure.

The processing operation should be analysed thoroughly and methodically. Provision should be made for all contingencies and answers foreseen down to the last detail. Nothing should be left to chance. Programming, i.e. the codification of the detailed sequence of instructions to be carried out, requires the programmer to be well acquainted with the survey to be processed and with the computer system he is operating (hardware and operating system). He must also have a perfect command of an advanced programming language.

Below are given the main phases of the processing of a survey by computer:

(a) Input: This is a relatively simple operation which consists of reading from the card file and recording on magnetic tapes, possibly with a preliminary sorting operation;

(b) Codification: This operation falls into two parts. It assigns to one or more inputs an abbreviated designation, INVTOT for example to represent total investment. It carries out automatic codification procedures which have been programmed, for example determining the code representing the employment group by comparing the total number of staff recorded with the category limits of set by the programme;

(c) Checking: This is the most complicated operation. The computer performs all the checking operations already done manually and (provided that the instructions have been programmed) automatically corrects omissions or anomalies. It also checks that the form and content of the inputs are in line with certain rules of computer technology. Checking is often done by the printing out of messages noting anomalies which must be eliminated before processing can continue;

(d) Processing: After a period of time which may vary according to the practice and experience of the operators, the checking operations are terminated and the card files are pronounced clean. Then the whole series of sorting, summarizing and calculating operations is carried out in accordance with the programmes - that is, with the risk that results fed out may be unintelligible because of a failure in the logic or instructions of the programmes;

(e) Editing: Although printers are now capable of printing 1,000 lines per minute, the results are not immediately transcribed on to paper because operations in the central processing unit are carried out at different rates and the printers function at different speeds. The results are sent via the central processing unit on a peripheral tape. A special programme, called the editing programme, is then required to transfer the results from the tape to the printer,

(f) Unified processing of data collected by several channels: This type of processing has its origins in statistical co-ordination: to avoid saturating sources, the collection of information for use by statistical and other administrations is organised in such a way that an enterprise is not asked to provide identical or similar data twice. The systems which have developed, generally called data banks, rely to a large extent on computers. Partial card indexes are formed by the administrations concerned in their own areas of specialization. The means of collating these partial card indexes are:

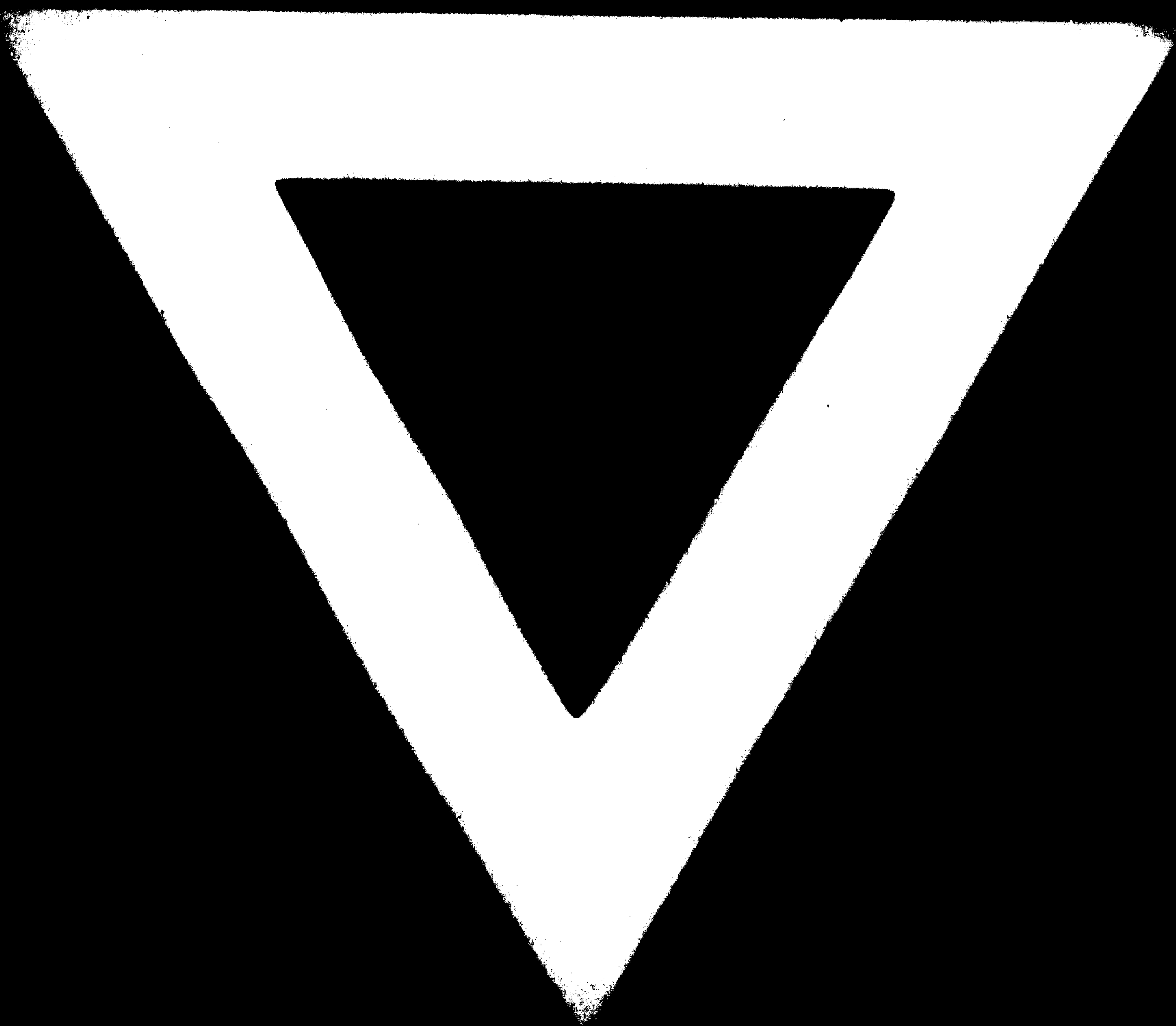
- The central identification register which assigns a particular identity number to an enterprise or establishment;
- A system of official nomenclatures and codes whose use is compulsory.

The highly complicated operations carried out by integrated processing techniques include the consolidation of the partial files into a unified form; collation, which identifies omissions or anomalies in the basic units seen as a whole; the aligning for each unit of information taken from different files; the compilation of the single register; processing, and editing.

Emphasis has been placed not only on the capacities of computer systems but also on the risks of failure. The person conducting the industrial survey, should, therefore, if he opts for this type of data processing, accept two constraints:

- An integrated conception of all the operations involved in the survey, instead of the traditional sequential approach whereby problems are solved as they arise;

- The testing of the processing system to be used for the survey by a trial run conducted before the survey itself is launched.



30. 11. 78