D00906

Distr.
LIMITED

ID/WG.[illegible]
[illegible] November 1969

ORIGINAL: ENGLISH

United Nations Industrial Development Organization

East African Working Party
on Industrial Programming Data

Nairobi, Kenya, 24 - 29 November 1969

THE INDUSTRY FILE SYSTEM: A DATA BANK FOR INDUSTRIAL PLANNING [1]

prepared by the secretariat of UNIDO

[1] This document has been reproduced without formal editing.

id.69-[illegible]

Table of Contents

INTRODUCTION

The starting point for the drafting of this report can be traced back to a proposal for a so-called Industry File System (IFS) as it was originally propounded at the First Session of the International Working Party on Industrial Programming Data which met in Vienna in November 1968.

At that session the IFS was proposed as a simplified core data bank that might permit the adaptive and flexible use of continually up-dated factory-directory type information as a supplement to regular statistical industrial surveys and registers. It was hoped that this core system would, despite its simplicity, be susceptible to versatile uses: that is, as an aid in market identification, inter-industrial programming, surveyance on project preparation and implementation, and a variety of other tasks which are generally associated with the term industrial planning.

If the original IFS proposal could be faulted on the ground that it merely indicated a theoretical possibility, it could also be said that it nonetheless served as a necessary first step toward the goal of creating an institutionally supported, versatile and - what is really important - readily implementable mechanism of data supply to meet the daily needs of those who are engaged in the various phases of industrial planning.

The original proposal for the IFS, as presented at the First Session of the International Working Party, was in a highly tentative and rudimentary form. The six-page note given in the Aide-mémoire (ID/WG.23/1)[2] was obviously concerned with what could be the point of departure for constructing a broader framework of IFS in its entirety. That is, the particular data filing scheme suggested at that time as the so-called "minimum" version, with a certain limited range of specifications in terms of coverage and items of information, is actually meant only to indicate the main elements to be considered for the core part of IFS, to which diverse types of data from a variety of sources should be linked in accordance with specific uses to be envisaged.

[2] UNIDO; "Aide-mémoire: purpose, scope, operational framework and provisional agenda for the First Session of the Working Party on Programming Data" (September 1968), especially Annex III, pp. 24 - 29.

This paper is thus intended to re-formulate the Industry File System as a general framework in which the linkage of various potentially useful records or data files is to be effected. The actual fillings to be put into such a general framework indeed depend on their designated purposes and the programme of implementation should be properly phased in time, according to the priorities to be agreed upon among participating institutions.

Among the possible priority fields of application of the IFS techniques, particular emphasis may be placed on the tasks related to industrial planning on the project level: i.e. the endless, iterative process of identifying and evaluating the shortcomings and the potentials of existing industries and planned investment projects. Attention will be paid further to the fact that this process of industrial project planning, participated in by numerous individuals and institutions, is not only the "user" of the outputs of the IFS, but also is the key "supplier" of the primary data on which depends the actual usefulness of the IFS operations for the process. The IFS proposal should thus not be conceived as merely a passive organ receiving and combining the existing scattered information. But the proposal should be considered, in its implementation phase, as including an active effort on the part of its users to link their routine tasks with the possibilities offered by such a data-processing facility.

To this end, the IFS proposal will have to be examined parallel to the substantive work involved at various stages of industrial project planning. Among others, attention is invited to the possibility for instituting a standardized system of recording specific project ideas, investment opportunities, feasibility studies, loan applications, investment licenses, etc., and developing analytical uses of such records linked to the information on existing factories and enterprises. As a first step for looking into this possibility, two additional working papers are being prepared: "Stages of project preparation introductory study" (contributed by D. Slimak) and "Data requirements for specified stages of industrial project preparation" (contributed by E.G. Bröder).

As regards the main technical aspects of the core part of IFS, the standard techniques and procedures developed in the form of Register of businesses need to be reviewed. For this purpose, the reader is referred to "The Register of Businesses - a first step in the development of an Industry File System" (By I. Osipov).

With the support of these papers dealing with the main technical details of the selected aspects of the IFS, this paper addresses itself to the task of drawing up the over-all framework of programme along the lines of the Industry File System. Part I provides a brief review of the background against which the idea has evolved. Part II is devoted to an exploration of how a data bank operation along the general lines of the IFS should best be structured and planned. Finally, Part III treats several of the central organizational and management problems which are likely to confront a project of this sort.

I. BACKGROUND

A. The proposal in its original form

The original proposal, put forward for the deliberation of the International Working Party on Industrial Programming Data at its first session (1968),[3] represented an idea that is yet at a very early stage of exploration. This first draft proposal, it is recalled, suffered certain defects.

The most serious of these is that it was consciously linked, rather excessively, to the anticipation that what would actually be feasible, in the face of the great expanse of needs, might only be precious little. It was thus concerned with what could be a minimum sensible system that would be more or less immediately implementable. As a result, the proposal lacked the indication of a broader framework in which any such minimum system was to be fitted and programmed for gradual extension.

For the reader's convenience, the main features of the originally proposed scheme are summarized as follows:

1. The IFS would cover both the information on actual industrial establishments and on potential industrial establishments (or investment projects);
2. To start with, the IFS would utilize the "directory" of factories or establishments to which only a few items of information would be linked. The information to be contained in this master file, so to speak, would be all readily available from the existing source, perhaps except one, i.e. listing of well-specified products for each industrial unit, with a classification system detailed enough to match the level of specification normally required for industrial planning on the individual project and commodity levels.[4]

[3] See ID/WG.23/1 "Aide-mémoire: purpose, scope, operational framework and provisional agenda" (September 1968), Annex III.

[4] The remaining six items were: (i) location of establishment; (ii) establishment status (independent enterprise, branch establishment, subsidiary, etc.); (iii) year of reference; (iv) year in which commercial production was first started or codified stages of project implementation; (v) total employment; and (vi) gross annual sales. However, no particular suggestions were made as to the methods by which these specific items were to be handled in a machine-readable form. Technically speaking, for example, item (ii) might be handled as an integral part of the system of establishment identification numbers; item (iv) would involve a technically slippery element; etc. These points are treated in some detail in I. Osipov, "The Register of Businesses" (ID/WG.52/4).

3. These items of information would be limited in their usefulness by themselves, but considered as being an absolute minimum in the sense that, were it not for even this much information, thoroughly up-dated, with good cross-sectional coverage, other potentially useful (and more or less readily available somewhere) kinds of information, if any, would remain footless or even mythical.

4. Leaving aside the possibilities for linking this minimum data file with other data from a variety of existing sources (which constitute actually the key to the "framework" of IFS in its entirety), it was hoped that even the minimum version, alone, would prove practically useful if established commonly in a group of countries with a mutually recognized interest in industrial co-operation.

B. Revision of the proposal - a broader frame of reference

The subsequent deliberations of the International Working Party on this tentative scheme made it clear that:

(i) There could not be any unique "minimum" system that would be equally feasible and useful for different countries; the real issue was rather "where to start" than "where to end"; and

(ii) The proposed scheme should thus be considered as suggesting simply a point of departure where a reasonably comprehensive inventory would be taken, on a continual basis, of the existing factories and note-worthy industrial projects.

With this re-interpretation, the Industry File System needs to be re-structured as a broader frame of reference for various specific programmes intended to exploit the possibilities offered by the techniques of "record linkage". From the standpoint of actual procedures of planning and implementation, the main points raised in the original version, as mentioned earlier, may be still relevant, but are obviously incomplete.

The specifics of possible implementation programmes of the IFS for various countries are yet to be studied by appropriate field teams. It is hoped that the discussions to be afforded by the East African Working Party on Industrial Programming Data will provide a further elaboration in technical and institutional terms of the particular frame of reference of IFS for such field studies, with due attention to the conditions prevailing in the several developing countries considered. At the present stage, it would suffice to describe the framework of IFS in the following general terms:

(i) The IFS will apply the modern "data bank" techniques to the problem of storage and retrieval of a broad range of data to be assembled from various (at present scattered) sources;

(ii) These sources include the various administrative records available in the existing governmental and non-governmental institutions, as well as the current national statistical programmes, and a programme for the IFS will necessarily include the establishment of appropriate agreements among the participating institutions;

(iii) Technically, the IFS will start with the Register of Businesses or a similar filing system concerning industrial activities, which will permit versatile linkages of the records from various sources in accordance with specified needs;

(iv) The Central Register of Businesses for the IFS or the core file (or files) will be established focussing upon individual industrial (or economic) activity units, both existing and planned;

(v) As regards the treatment of planned activity units (investment potentials and projects), the IFS should preferably include an active supporting programme that is aimed at generating a new flow of information concerning industrial project planning and implementation, from one development institution to another, channelled through the given data bank; such a supporting programme will in turn include the establishment of a set of standardized forms to be commonly used by all participating institutions for regular reporting purposes, and also a main programme of analysis to be applied to the reported project data for consistent evaluation and planning of project activities;

(vi) It is not pre-supposed but is recommended that the IFS be structured in a basically comparable or inter-linkable manner for two or more (neighbouring) countries which have a recognized potential for industrial plan harmonization and co-operation.

II. THE PLANNING OF AN INDUSTRY FILE SYSTEM

Given the over-all frame of reference for the IFS as above, it is clear that any attempt to specify the details of the system must begin by focussing on the data which are to provide its backbone. To this end, it will be recalled that it was proposed that the IFS be developed from start to finish as a "user-oriented" system, and that the users in question are those associated with industrial development planning.

An introductory discussion of both "industrial planning" and the "data prerequisites for industrial planning" is to be found in the document

"Agenda and background information" (ID/WG.52/1) which was prepared by UNIDO for the East African Working Party on Industrial Programming Data. A far more detailed discussion of these issues is provided by the UNIDO study entitled "Data requirements for industry analysis and programming" (ID/WG.23/4). A careful review of these two documents should serve to explain the central characteristics of the individual pieces of information which are to enter the IFS, at one stage or another, and in one form or another.

Apart from the general review of the data requirements for industrial planning, it is obvious that more detailed and selective specification should take place only when the specific context within which the System is to become operational has been determined. This task is to be fulfilled in the post-session phase of the East African Working Party. At this stage, it would be rather more important to establish the general guidelines for structuring a programme of implementation for the IFS.

A. General guidelines for the development of the System

The considerations listed below have the character of general guidelines only; they are not meant to provide a manual or handbook on "how to develop a data bank". Detailed instructions of the latter sort could - and should - be prepared in each instance where the decision has been made to implement an action in this direction, and possibly after having conducted a pre-test and pilot study in the specific operation context within which an IFS is to be established.

(i) Capacity: A decision to build up a data bank does not mean that it, once begun, will work at full capacity from the first moment. Experience shows that it is desirable to develop a data bank gradually, beginning in a modest way and with time to widen its scope regarding the number of units to be included and to deepen its contents with regard to the number of pieces of data and different records to be linked together, the scope of the routine processing programmes to be run, etc. A gradual setting in motion of the data bank will not only allow (a) potential users of the data to re-examine their requirements in the light of the data which are received, to develop their own methodologies for better-informed work, and to alter their demands accordingly, but also (b) time for the data bank managers to learn from their experience about the problems which arise during collection, processing, and dissemination of the data, and to improve their work concurrently.

(ii) <u>Data suppliers = data users</u>: A data bank along the lines of the IFS can be built up only on the basis of close collaboration between the managers of the data bank and the suppliers of data. It is desirable - and it should be properly incorporated into the IFS programme itself - that the suppliers of the data be also those who will profit most from the existence of the data bank. Only if the suppliers of data have a genuine stake in the success of the data bank can it be expected to garner the support necessary to guarantee its success.

Moreover, it is strategically important that the users of the data bank are able to receive the output as soon as possible once the data bank begins to function. Thus, it is advisable to start record linkages at first by concentrating on concrete orders received by (or anticipated from) the users. It will also be necessary to pre-specify forms of dissemination, which are indeed different from the conventional statistical publications.

(iii) <u>Location of data bank vis-à-vis other government units</u>: It may be considered generally advisable to organize the data bank as an independent unit outside of the statistical office, because the activity characteristics of the data bank will basically be quite different from, and in some ways may involve even statutory contradictions to, those of the national statistical office.

For example, it is known that the activities of a statistical office are based on the principle of the <u>confidentiality</u>; this principle is considered essential for the effective operation of the statistical office. Moreover, the standard procedure of statistical offices is to produce summarised data only and to publish the results in the form of aggregated statistical tables.

On the other hand, the data bank will freely utilize the data derived from administrative records (which are mostly non-confidential); these data will be retained at the level of the individual units to permit versatile uses by sorting, aggregation, cross-reference, etc. geared to specific uses. The data obtained under the Statistical Ordinance by the statistical office may (and should) be linked to the data bank only upon explicit consent of the businesses.

Thus, any attempt to combine the activities of the two may create difficulties, undermining the alleged dependence of data reporters on the confidentiality of the information they make available to the statistical office, and on the other hand, to create additional obstacles which might hinder the effective operation of the data bank.

(iv) Relations with statistical offices: The foregoing brings up a special and very important problem which is posed on the mutual relations between the bank and the various statistical offices. This is important because the confidentiality principle has often served in the past to defeat similar data-centralization projects. It must be recognized by all parties, however, that a great deal of needs and possibilities exist for collaboration between the various statistical offices and the data bank. Some of these are given below:

- In many countries, only some of the data obtained by the statistical office is confidential. This might allow certain pieces of the remaining data to be fed into the data bank. In Uganda, for instance, it is permitted to publish figures for one or two establishments relating to such items as employment and production.
- In some countries, certain pieces of data from the Register of Businesses (namely the address, activity, legal organization, connexion and kind of establishment) are held to be public property and freely communicable to anyone.
- In other countries, it is a standard practice of the statistical office to ask in questionnaires if the informant agrees that certain pieces of data are not treated as confidential, and hence can be circulated in accordance with specified limits.
- There are also, in many cases, provisions which permit institutions requiring statistical data to ask the direct permission of the supplier if the statistical office can "free" certain data for their use. In such cases either a "blanket" permission may be granted, or the request may be repeated in each specific instance. It has been observed that, to avoid the burden of duplicative multiple enquiries, blanket permission is often granted by the respondent establishments for all other than profit and certain other key pieces of operational data.
- Attention should be paid to a special group of data collected by the statistical offices from ordinary administrative sources. Often, as with the Greek King Midas who turned everything he touched into pure gold, that such data - once routinely available - becomes, on being "touched" by a statistical office, "confidential". This is a situation which should and could be overcome.

Finally, it is hoped that the statistical offices, realizing that they too can profit from an effective data banking system, will be willing to co-operate in the task of building up an institution which should be a major complimentary tool for fulfilling data needs.

(v) Dynamic context of a data bank: Once the decision has been made to launch a data bank, it must from the first be taken into account that it is destined to operate in a fully dynamic environment. The conditions to which it will have to adapt will be continually changing.

On the one hand, the over-all demand for data will grow rapidly mainly as a result of the increasing trend toward data-intensive planning techniques, as well as the gradual expansion of the data requirements for a variety of administrative procedures. In addition, it can be anticipated that the data bank will be faced with increasing demand for new kinds of data: more specialized data, up-to-date data, data readily arranged for specific uses, etc.

But the dynamic elements which provide the context into which any data bank must fit are not limited solely to the demand side. They relate also to changes in the conditions for reporting information, changes in the techniques of data processing, and changes in the methodologies for approximation.[5]

(vi) Effective "balance" of work load: In drawing up a detailed programme for the development of a data bank, special attention should be paid to the distribution of the total work load between the data bank and the other institutions involved. This is mainly a matter of apportioning the work load appropriately among data reporters, administrative authorities, and among the consumers. If a data bank is to be practically implementable, maximally useful, and thoroughly cost-conscious, it is important that its organisers do not attempt to absorb all of the costs and operations that will spring up in its wake. Any attempt to do so might well spell the death of the project.

(vii) Development programme for data bank: It is essential to the success of the operation that a detailed time-phased programme be drawn up

[5] See, e.g. Problems of organizing a modern statistical service, Conference of European Statisticians/265, 1968.

for the development and operation of the bank, estimating the resources required and the expected costs involved at each step, and determining which project activities are to be handled both within and outside the bank.

5. <u>Preliminary outline for an IFS</u>

Given below is a preliminary outline for the development of an IFS. It begins with a brief consideration of the general characteristics of the proposed data contents of the system. This outline is intended to serve primarily as an orientation scheme. A precise statement of the required data inputs can be given only when the concrete demand and data sources are known in a given country or region.

In planning such a review, the projected demand for and the existing sources of data must be examined with respect to the following:

(i) Type of information
(ii) Scope
(iii) Specific data categories
(iv) Classification system to be used
(v) System of measurement
(vi) Calculations
(vii) Periodicity
(viii) Storage system
(ix) Institutional source
(x) Reliability.

<u>(i) Type of information</u>

The types of data which will go into an IFS data bank may be divided into three main groups:

(a) <u>Output data</u>: production, shipments, sales, value added, etc.
(b) <u>Input data</u>: employment, hours worked, materials consumed, machinery
(c) <u>Other data</u>: capacity, installed machinery, power equipment, orders, investments, taxes paid, loans received, etc.

<u>(ii) Scope</u>

The scope of the data to be plugged into the system must be defined mainly with regard to the following: location, industry branch, size of unit, form of ownership, status of implementation (especially for new units to be constructed), etc.

(a) Location: It must be specified whether data is required on businesses which are urban, rural, in all or in part of the given country.

(b) Branch: It must be specified if the bank is to cover, for example, Mining and Quarrying, or Manufacturing, or Construction, or all, or only part of one of these.

(c) Size of unit: A floor (and/or a ceiling) for the size of units to be covered must be defined in terms of the number of persons employed, gross revenue per annum, value added, investment, or some other such determinant characteristic.

(d) Ownership: A decision must also be taken as to the scope of the bank with regard to different types of ownership to be considered, such as: single-owner firms, private partnerships, private limited companies, public limited companies, etc; or with regard to different forms of ownership, e.g. private, public stock, co-operative, parastatal, national, etc.

(e) "Planned" units: It is recommended that not only should the IFS contain data on existing units, but that it have a register for "planned" units in various stages of the process of being established. The information on such units must be identified with reference to its respective stage of preparation, from the pre-investment study through the pilot production phase. The usefulness of the data bank for industrial planning purposes would very much depend on the scope of the bank in this respect. As mentioned earlier, this calls for active programming and management on the part of the supply sources of the information on planned business units - the development institutions, themselves, that participate in the IFS programme.

(iii) Specific data categories

It is important that full consideration be given to the various categories of data at the unit level in making a final decision as to the scope of the IFS, such as the following:

Category	Characteristics
Production	- detailed quantity/value breakdown by major products
Sales	- detailed breakdown of marketing channels (for domestic and export markets)
Employment	- detailed breakdown by sex, age, occupation, employment status, etc.
Machines installed	- detailed breakdown of source (local production or imports)
Orders	- new, booked, and unfilled
Inventory	- production materials, work-in-process, finished goods.

The breakdown of production by individual products is of great importance for the users of the [illegible]. It is tentatively recommended to use the "Classification of Commodities by Industrial Origin", according to which for each ISIC group is listed all items of commodities belonging to that group. The United Nations Statistical Papers, Series M, No. 4, Rev. 1, Add. 1,offers a basis for this approach. In 1968, a new, Rev. 2 edition of ISIC was brought out by the United Nations, and it is anticipated that a new, up-to-date classification of commodities by industrial origin will become available in the future. To this end, it should be mentioned that there are often cases when an establishment classified in one branch, also produces some products to be listed in other branches.

(iv) Classifications

It is critical to specify which system of classifications (international or local) is to be used (or is required) for grouping the data, e.g. Standard International Trade Classification (S.I.T.C. Revised), The Brussels International Nomenclature (B.T.N.), International Standard Industrial Classification (I.S.I.C.), International Standard Classification of Occupations (prepared by International Labour Office), to name but a few. Moreover it will be advisable to make use of "cross-classification keys" for translating the local branch classifications of each of the participating countries into I.S.I.C. Rev. 2 (1968), or any other suitable classification system.

For some purposes, the classification of commodity groups by industrial use, rather than by origin, would be found to be useful. For example, the data for export projections and for local consumption can be treated only on the basis of consumption in the importing countries and on the basis of intermediate and final uses in the local market. For some other purposes, the classification of commodities by types of production process would be desirable.

A system of classification normally involves an element of compromise between different (sometimes mutually exclusive) criteria. The classification of commodities for planning purposes normally relates to commodity groups only. A more specified classification, including individual products determined by specific characteristics, e.g. colour, size, etc. may be available in the price catalogue or prospectus of the enterprises.

Such detailed classification is seldom needed for general planning purposes, except for certain particular commodities.

To ensure a reasonable degree of flexibility (to enable the addition of new products; the integration and differentiation of groups of commodities according to specified needs, and to facilitate the data processing by computer), it is advisable to use the decimal systems of classification code for main commodity groups. The serial system may be employed to cover the variations within each well-defined commodity group.

Of great importance for planning purposes is that not only the products but also the materials are included in the classification, since the products of one branch of the economy are used in another branch of the economy as materials or intermediate inputs; the identical name of a commodity, no matter if referred to as product or as material, should be used in order to enable the analysis of input-output relationships in an economy.

(v) Systems of measurement

Uniform units of measurement must be defined for each unit of data, e.g.:

- products, by quantity units (tons, m^3, number of units, etc.);
- sales, by value (at current prices or at constant prices);
- gross output, by value (at market prices or at factor cost);
- inventory, by value (LIFO or FIFO); etc.

(vi) Calculation

The data required must also be defined according to the extent and type of calculation involved, e.g.: individual, aggregative, cumulative, derivative, and so on.

(vii) Periodicity

One must define, in advance, the frequency with which data is supplied and the timing of reporting for each source.

(viii) Storage

During the review of the information sources, the system to be used for storing the data, e.g. manually written reports (list), punched card, magnetic tapes, etc. must be specified.

(ix) Source of data

The main sources of data for the IFS must be determined among such sources as: statistical reports; administrative materials registered to satisfy direct administrative requirements; mechanical counters (e.g. traffic counts); direct enumeration; transactions data, etc.

(x) Reliability

The reliability of the data must be defined by the completeness of coverage of the surveyed population, the reliability of the responses, estimated sampling errors, and so on.

* * * * *

After a detailed comparison between the information to be requested on a priority basis and the existing sources of relevant primary data, it should be possible to evaluate the need for the exploitation of additional data sources. The need for additional data should be weighed in relation to the price to be paid for obtaining them. A final decision can be taken as to the scope and contents of the additional data to be entered in the IFS data bank, only after such detailed evaluation has been carried out.

III. SOME TECHNICAL AND ORGANIZATIONAL PROBLEMS OF A DATA BANK

By way of providing general guidelines for planning further the early phases of operation of a data bank, several main aspects of data bank operations are sketched out in the following.

A. Information inflow

The data flowing into the data bank have to undergo a complex process of adaptation in order that the record linkage phase may be implemented. The steps of this process are quite similar to the normal statistical office routines dealing with ordinary survey data and it is not necessary to explain it in detail here.

Main steps for handling the in-flowing data may be stated as follows:

- Assembly;
- Manual checking and corrections;
- Coding;
- Punching;
- Mechancial checking and corrections;
- Processing;
- Tabulation;
- Final checking;
- Analysis;
- Printing out the results as required.

The scope of each stage and the depth of the treatment to be accorded to it have to be decided in accordance with the needs of the data bank.

The manifold forms in which data are stored in those institutions supplying the primary data demand painstaking preparatory work of standardization and uniformalization for machine processing. This preparatory work is essential, since data of the same subjects flowing in from different sources often differ in the way they were prepared or even in the underlying definition of concepts.

B. Record linkage

The techniques of record linkage can be illustrated by means of a three-dimensional space, the dimensions being the statistical units, the characteristics of the units and time of reference. The simplest types of record linkage consist of the collation of data records along only one of the three axes, e.g.:

(i) Along the axis of statistical unit - collating data of related individuals regarding a characteristic at a given point of time (e.g. linking the data from different branch establishments of the same enterprise);

(ii) Along the axis of characteristics - collating data relating to different characteristics of each given unit at a given time point (e.g. linking the data on sales obtained from sales tax records with the data on employment from a census questionnaire for a given establishment for a given year);

(iii) Along the time axis - collating the data relating to a characteristic of the same statistical unit at different points of time (e.g. data from the monthly reports of the employer to a national insurance corporation).

In general, however, record linkage can be performed by collating data along two axes simultaneously.

Record linkage is usually required for collating data from different administrative and statistical sources, recorded at different times and by different methods. The basic process is that of matching two sets of records, each of which refers to a certain population of statistical units. The relationship between these two populations can be one of the following:

- The statistical units theoretically represented by the two sets of records are in one-to-one correspondence;
- One population of statistical units may be a sub-set of the other;
- Neither one of the population of statistical units is a sub-set of the other.

In general, the matching is designed to locate those records which represent units common to both populations and to pair them correctly in order to use the combined data. (In some cases, the non-linked records can be of interest, either as a by-product of such linkage or for the very purpose of detecting such records.)[6]

The institutions which possess data to be processed by the data bank will constitute important "members" of the data bank. They will probably

[6] Bachi, R., Baron, R., Nathan, G., "Methods of record-linkage and applications in Israel", Proceedings at the 36th Session of the International Statistical Institutions (Sydney), 1969, pp. 766-784.

profit from the data bank. It is thus preferable to start the record linkage with those institutions.

Before delimiting the operation of the data bank in its full scope, it would be desirable to actually carry out several programmes of record linkage so that the debugging of these programmes can contribute to the timely initiation of an action for the main lines of data supply and utilization, both within and outside the data bank.

The record linkage can be carried out either:

(i) On an ad hoc (single instance) basis to meet a special need; or

(ii) On a recurrent basis using the regular inflow of data for a range of pre-specified needs.

The single-instance record linkage takes place to meet special research requirements of the data bank members. Suppose, for example, that a water authority supplies yearly to the data bank data on water consumption of individual industrial establishments. This data can be combined with the information received on the output of these establishments, so that the changing relationship may be obtained between water consumption and industrial production over different years. This kind of linkage would facilitate the estimation of future water consumption to be envisaged under given industrial development programme for the coming years.

A similar analysis can be performed with regard to electricity consumption in Kwh of industrial establishments. Incidentally, the data available from annual statistical surveys on electricity consumption may be given in value terms only. Such value figures might include different categories of electricity consumption to which different tariffs apply. They might even include expenditures on the installation of electrical transmission lines and their repairs. In estimating the real consumption of electricity for industrial production, the administrative records of an electricity corporation, linked to the reported data from industrial establishments, may offer an improved basis.

In some countries, new foreign enterprises may be given preferential treatment in the form of grants, loans, lesser taxes, exemption from customs duties on imported equipment, etc.[1] It would be of great interest to compare the development of these enterprises with the development of enterprises which did not enjoy such advantages. For that purpose, record linkage can be performed for the records of preferments granted as available from the administrative files of government offices, banks, etc. with the records of the enterprises obtained by the bureau of statistics concerning employment, wages, investments, input, output, etc.

Recurrent programmes of record linkage are to be established, most typically, to link the relatively static characteristics of business units (size, industry branch, location, etc.) to the data on changing activities of the units (production, imports, investments, etc.).

For example, figures on imports of investment goods obtained monthly from the customs authorities can be combined with data from the Central Register of Businesses on the particular branch or product group in which the importer is dealing. Thus, this enables an analysis of the imports of investment goods by branch of destination. Also, by combining these data with the available data on local production of investment goods (machinery, equipment, transport vehicles, etc.) one can get current (monthly) indicators of investments in the various branches of economy.

It is sometimes difficult to trace the final use by industries of the commodities which are imported by trading companies and not by industrial users themselves. Up-to-date information of this kind would be useful for development planning purposes. When the linkage of available records does not alone produce satisfactory results, a case study may have to be carried out once in a while, seeking the keys by which to approximate the distribution of imports by industrial uses, so that meaningful results can be obtained on a recurrent basis.

C. Information output

As discussed above, the output of single-instance record linkage and that of recurrent record linkage need a different treatment. The non-recurrent type demands individual programming in each case, while the

[1] e.g. According to the Foreign Investments Protection Act (1964) in Kenya, the owner of an Approved Status Certificate is entitled to repatriate earnings and capital, to receive duty refunds on imported materials, investment allowances, financial participation, etc. (Development Plan 1966-70, 1966, p. 238.)

recurrent type is susceptible to routine procedures for manual and machine treatment, which reduce the processing costs.

Programming costs constitute a significant proportion of the total costs of EDP. It would therefore be profitable to put some programmes or some part of each programme into a general form that can be used for certain typical steps involved in most of the different processing tasks.

The individual steps in the process of information inflow have to be programmed in detail, specifying the particular machine, form, and printing techniques to be employed. Mechanical quality control should be gradually introduced and the necessary publications be prepared directly by computer. For this purpose, rationalization of manual routines and of form designs should be accomplished with regard to all steps including reporting, checking, coding and punching work.

A careful analysis of the process of information inflow and outflow should be conducted to ascertain what kind of logical checking will be successful and what kind can be left out. The printing method has to be programmed to ensure that the right form is achieved to prepare directly photo-copies and printings from the computer. If the data have to be re-written for printing purposes, another source of mistakes will open, and another checking period will be needed. Routines have to be established for regular checking work on the entire system and for follow-up at a later stage on the result of the work carried out.

It is only some time after the bank started actual operation that the system will be properly working. The preparation of a detailed operation manual can then be started.

Programming with regard to the working relationship between the data bank and its member institutions will receive no less attention than programming of the internal mechanism of the bank. For one thing, it is important to have well-designed forms to facilitate the supply of needed data from these institutions. For another thing, it is equally crucial to establish a sensible programme of dissemination of the output tailored for the regular use of each participating institution. Every effort must be made to produce actually useful results, without delays, if only in a limited, rudimentary form to start with.

D. Management of data bank

It is necessary to form an inter-institutional organ to manage the data bank. In some cases, such an organ may be best put under the direction of an inter-ministry commission, which may be joined also by other users and suppliers in the private sector. Such a commission ought to be organized with a dynamic orientation. It can be enlarged to cover the interests of a growing number of participants in the data bank. The commission will play a crucial role in providing the necessary social and institutional backings for the activities of the data bank, including both its users and suppliers.

As mentioned above, the activity of the data bank begins with the establishment and organization of a register of businesses. The core of the data bank, thus established, will be gradually enlarged as a growing number of data sources is linked to it.

It is recommended that the unit handling with the register of businesses at an early stage consist of two sections:

(i) One section dealing with the compilation and bringing up to date of lists of employers in all branches of the economy; and

(ii) The other dealing with the preparation of area samples for the purpose of obtaining estimates of self-employed and other businesses not covered by the lists of employers.

The first stage of setting into motion a register of businesses needs a number of persons dealing with the preparation of data for machine processing, manual checking and other operations. For a register of, for example, 25,000 establishments, the necessary personnel for the first year may be the following:

(a) 1 person with advanced training in statistics and economics;

(b) 6 - 8 persons with a medium degree of education in social science; and

(c) Several persons to be trained for technical jobs of limited scope.

Once the register of businesses has been established, its maintenance and up-dating, except its extension of significant scope, will require no more than 1 skilled and 3 unskilled persons.

E. Problem of confidentiality

The data to flow to the bank from different sources varies in the degree of confidentiality. Some data will come from lists that are open to the public, some from lists intended for given official use only, and some from the statistical offices which are subject to restrictive disclosure rules. Secret data integrated with non-secret data results in altogether secret data, and must be treated in accordance with the same rules that apply to the (originally) secret part of the data. This, of course, would depend on the manner in which the integration of the two types of data was performed.

In handling the information inflow, the data bank should establish clearly differential treatments of confidential data files and non-confidential data files. In many instances, the records on individual statistical units obtained by national statistical offices under given Statistical Ordinance must be kept in confidential files and handled accordingly. Administrative data files, containing information regarding investment licenses, import licenses, construction awards, public loans, trade union reports, etc., may be handled mostly as non-confidential.

The linking of administrative data files with the census returns of the statistical offices will be one of the most important possibilities that have to be thoroughly explored in determining the exact scope of the data bank. When the data bank is instituted outside a national statistical office, as suggested earlier (see discussions in Part II, A), the statistical ordinance is likely to preclude the possibility for placing the census data file, as such, at the disposal of the data bank. Solutions should then be sought

- by establishing an appropriate EDP capacity within the statistical office, whose outputs will be programmed in concordance with the confidentiality rules, but in a manner readily linkable to the data processing programmes envisaged in the data bank, and at the same time,
- by obtaining written consent from each particular business of interest for the use of particular items of information for specified purposes connected with the data bank operations.

The latter procedure is particularly important and the experience indicates that it actually offers enormous possibilities for circumventing the legal restrictions concerning the use of data from the statistical offices. Such consent from the businesses should be obtained, however, with carefully formulated criteria regarding the selection of relevant categories of information and ways in which they are combined with other categories. In this respect, particular attention is invited to the following paragraph of the Statistical Ordinance under which the East African Statistical Department:

> "Provided that nothing in this section shall prevent or restrict the publication of any such report, abstract or document without such consent where the particulars therein render identification possible merely by reason of the fact that they relate to a business or undertaking which is the only business or undertaking within its particular sphere of activities, so, however, that in no case shall such particulars render possible identification of the costs of production of, or the capital employed or profits arising in, such business or undertaking."
>
> (Nairobi, 4 July 1961)

Even where individual items of data relating to a given business are not secret, their integration may result in an information that might be harmful to the business should it become public knowledge and reach the ears of its competitors, suppliers, customers, banks dealing with it, etc. For this reason, there is always obligation on the part of the authority administering the data bank to see to it that detailed data on a specific business that which permits its identification will not be released unless that business's explicit permission is obtained.

There are several studies devoted to this aspect of the data bank, all emphasizing the need for constant awareness of the ethical problem connected with the operation of a data bank.

[redacted]

74.10.2