also requires contacts with professional activities and with *developments in methods* that occur there. The instruments available for the production of statistics are continually being improved, and this improvement should gradually affect the performance of the various functions specified in the diagram. The following have been indicated, though the list makes no claim to be complete:

A. statistical methods
B. equipment techniques
C. organization and systems analysis
D. design of forms
E. printing techniques
F. "public relations" activity.

In the case of activities A—D there is a mutual influence, as the CBS contributes to the development in these spheres by its internal activities.

On statistical file systems II[1]

by Svein Nordbotten[2]

0. Introduction

At the Nordic statistical meeting in Helsinki 1960, some ideas were presented under the name *Statistical file system* (2). The aim of the present paper is to discuss some further views connected to such a system and report how we are implementing some of these ideas in Norway. The presentation is to a large extent based on impulses from discussions with a number of statisticians.

The name statistical file system may give the impression that there is something quite new we are going to discuss. This is not the case. The system to be presented is to a very large extent based on old ideas which the technical development has now made possible to realize (3).

[1] Translation of a paper presented for the Nordic directors of statistics in Copenhagen, June 16—18, 1966.
[2] Central Bureau of Statistics of Norway, Oslo.

1. Needs for statistical information files

The basic idea behind the statistical file system is that any new observation represents an increment to the general knowledge and has a value of information if it can be made available for practical utilization. Previously the cost of making stored observation available was often larger than its value of information. The ratio between cost and value is now decreasing both because the storage cost per observation is being reduced and because the linkage possibilities are increasing which also increases the value of each observation. Finally, the value of information is growing because the demand for statistics is increasing in general.

The demand for statistical information is partly expressed as requests for special processing in addition to what can be presented in standard sta-

tistical volumes. Frequently the demanded processing requires that data from the same units, but collected from different sources or at different times, must be linked together.

The scientists and research workers are not quite satisfied any more by having to work only with aggregates for groups of primary units because these aggregates can give knowledge about the variation between groups while the variation within the groups is lost, and because a number of models concerning the technical or behaviour functions of the individual units cannot be aggregated to corresponding macro functions. The alternative to macro-computations based on aggregated data may therefore be micro-computations followed by an aggregation of the results.

To satisfy such needs will require extensive information files in which data from different sources are compiled, systematisized and stored in an easy accessible way.

Modern data processing equipment may contribute to the establishment of such a readiness in two ways. First, they represent efficient filing tools which make it possible to file very large masses of data automatically. Second, the application of data processing equipment in administrative agencies will make data originally collected for non-statistical purposes available for statistical use at a low cost.

Similar views have also been expressed in other countries. In Sweden a committee for development of a statistical file system has been established. In USA a committee, the Ruggles committee, sponsored by the Social science research council, emphasized recently the need for a national data center for systematic preservation of data collected by different agencies. The idea has been studied by the US Bureau of the budget and the result presented in a detailed report on the needs for such a data center (1).

In the following sections, the principle views and practical approaches followed in Norway are presented. The coordinating and planning required is done within a continuous planning and progress reporting system.

2. Theoretical considerations

It may be useful to start the discussion with some theoretical considerations. We assume that the size of the national product is partly depending on the knowledge incorporated in the society. Knowledge is not consumed, but is made use of in processes more like the way production capital is used in a production process. A part of the knowledge is of statistical nature, i.e. it concerns groups of units or phenomena, not individual units. We shall here study the changes in statistical knowledge.

The supply of statistical information to the society takes place by the publication of results from statistical computations. Let this process be denoted by:

$$I = I(M,S) \tag{1}$$

where I is the supply of information per time unit, S is the stock of computed statistics and M a publication factor which expresses the utilization of S

tistical volumes. Frequently the demanded processing requires that data from the same units, but collected from different sources or at different times, must be linked together.

The scientists and research workers are not quite satisfied any more by having to work only with aggregates for groups of primary units because these aggregates can give knowledge about the variation between groups while the variation within the groups is lost, and because a number of models concerning the technical or behaviour functions of the individual units cannot be aggregated to corresponding macro functions. The alternative to macro-computations based on aggregated data may therefore be micro-computations followed by an aggregation of the results.

To satisfy such needs will require extensive information files in which data from different sources are compiled, systematisized and stored in an easy accessible way.

Modern data processing equipment may contribute to the establishment of such a readiness in two ways. First, they represent efficient filing tools which make it possible to file very large masses of data automatically. Second, the application of data processing equipment in administrative agencies will make data originally collected for non-statistical purposes available for statistical use at a low cost.

Similar views have also been expressed in other countries. In Sweden a committee for development of a statistical file system has been established. In USA a committee, the Ruggles committee, sponsored by the Social science research council, emphasized recently the need for a national data center for systematic preservation of data collected by different agencies. The idea has been studied by the US Bureau of the budget and the result presented in a detailed report on the needs for such a data center (1).

In the following sections, the principle views and practical approaches followed in Norway are presented. The coordinating and planning required is done within a continuous planning and progress reporting system.

2. Theoretical considerations

It may be useful to start the discussion with some theoretical considerations. We assume that the size of the national product is partly depending on the knowledge incorporated in the society. Knowledge is not consumed, but is made use of in processes more like the way production capital is used in a production process. A part of the knowledge is of statistical nature, i.e. it concerns groups of units or phenomena, not individual units. We shall here study the changes in statistical knowledge.

The supply of statistical information to the society takes place by the publication of results from statistical computations. Let this process be denoted by:

$$(1) \qquad I = I(M,S)$$

where I is the supply of information per time unit, S is the stock of computed statistics and M a publication factor which expresses the utilization of S

and the size of the publication issue. We shall call *S the capital of statistics* because it participates in the information process just in the same way as real capital in a physical production process, and may be used repeatedly. Statistical information may therefore be supplied without requiring any change in the capital of statistics by issuing new editions of statistical publications.

The capital of statistics is produced by computation partly on data for individual units and partly on previously computed statistics. The capital of statistics consists of aggregates such as totals, averages, etc. for groups of two or more statistical units. *Investing* in capital of statistics is done by computation and the investment per time unit is denoted by:

$$\dot{S} = S\,(V,U,S,D) \tag{2}$$

and defined as

$$\frac{dS}{dt} = \dot{S} \tag{3}$$

D is the stock of individual data which is called the *data capital*, V and U indicate the degree of utilization of S and D, respectively. Statistics may therefore be computed without any changes in D, e.g. by utilizing old data in computation of new statistics. In contrast to the capital of statistics, the data capital consists of data for individual, statistical units.

Investment in data capital, defined as

$$\frac{dD}{dt} = \dot{D} \tag{4}$$

is done by data collection.

To each of the above processes, $\dot{I}$, $\dot{S}$ and $\dot{D}$, and to the storage of S and D, costs are associated. The cost per time unit is expressed by

$$C = C\,(I,S,D,\dot{S},\dot{D}) \tag{5}$$

All variables in the system (1) — (5) are time functions.

The objective of the statistical system may be formulated as finding those time functions for $\dot{D}$, V, U, and M which, subject to the initial values for S and D, and the conditions (1) — (5), maximize a functional:

$$W = W\left(\left[I\right]_O^T \left[C\right]_O^T\right) \tag{6}$$

This is a typical problem of dynamic programming for a period of T units.

In a more elaborated model, different types and age classes of S and D with different costs and productivity have to be distinguished. In some cases, the productivity of certain types of statistics or data is so low that it has to be discarded, and the concept of depreciation may be introduced.

The difference between the statistical file system as outlined above and the more traditional way of describing a statistical system, is the introduction of the capital of statistics and the data capital. If these variables are substituted by their derivatives, we obtain formally a model representing the more traditional way of considering a statistical system the management of which becomes a static problem. In other words, within the system discussed here, the status at any time will be a function of previous development of S and D because computed statistics and collected data are considered reuseable. Within a system which does

not allow for repeated use of computed statistics and collected data, the situation at any point of time will be independent of previously computed statistics and collected data.

The above considerations may have implications for data collection, computation of statistics and publication of information. So far, the plans for data collection have usually been designed for one subsequent processing while the dynamic considerations above require that also probable, future computations in which old data are linked to new data must be taken into consideration. This may lead to plans for more continuous collection of data than the present. Plans for computation of statistics may have to take into account that time series for individual units are available which may be reflected in development and application of new estimation techniques.

The data capital and the capital of statistics are essential concepts in the system. The condition is, however, that they represent data and statistics which are organized in such a way that they can be utilized. This condition may be illustrated by a *data-box*, containing a number of small rooms for storing data. Each data is identified by the statistical unit to which it is associated and which has its permanent position along the first axis of the box, by the characteristic observed which has its permanent position along the second axis and finally by the period or point of time which has permanent position along the third axis of the box. The content of the rooms in a slice of the box across the time axis will give a data picture of the situation at a point or in a period of time. A slice across the axis of characteristics will represent a certain aspect of development, while a slice across the axis of units will tell the registered life story of a unit. A similar box may also be associated with the capital of statistics with the difference that instead of units each statistical group will have their permanent position along the first axis.

The conditions for the data box organization are that we have a system of permanent unit identification and standard codes for all characteristics. Within the capital of statistics, the national accounts system is an example of identifications and codes which make the statistics consistent and comparable over time. For the data capital similar systems have not been developed to the same extent.

3. Central registers

3.1. General considerations

A *register* is a list of the statistical units within a mass. The establishment of a register assumes that the definitions of the unit and the mass are determined. Both definitions lead to series of problems which are outside the context of this paper. The register represents a cross-reference between two types of identifiers, the *internal* and the *external* identifiers. The internal identifier which goes with the data in and out of the data capital a large number of times, should be constructed to occupy a minimum of space. In other words, the identification system should be as compact as possible to avoid that the data box becomes unnecessarily

wide because of unutilized positions along the first axis. In this way the needs for capacity and operation times may be reduced to a minimum. The internal identifier may be a number or a combination of characters. The external identifier is used to find the statistical unit when data are needed directly from the unit. It may be a name, an address, etc. In some cases the same identifier may be used both internally and externally, but usually the external identifier will be useless for internal identification because it is unnecessarily long and unstable.

The requirement to permanent identificators implies that definitions for *birth, migration* and *death* have to be specified for the different types of units. This leads to problems which were not always considered previously. Now they have to be straightened out if the registers are to be maintained and the condition of permanent identifiers satisfied.

It is of great importance that as many as possible of those agencies which collect data use the same registers. The data may then easily be transferred to the statistical system and preserved in the data box. It seems to be appropriate to charge the statistical system, which always will have the most complete data for maintenance, with the responsibility for such central registers.

In Norway the Central Bureau of Statistics keeps several central registers which are developed to satisfy the requirement of a statistical file system.

3.2. Central population register

The population registration in Norway is authorized by law of November 15, 1946, and the Central Bureau of Statistics has been charged with the responsibility for acting as a central office for the registration. To-day there are 460 local registration offices in Norway. In addition to the register documents required by the law there are punched card registers at each office. The registers serve a number of administrative purposes as well as being the source for the current population statistics.

In connection with the 1960 Census of population, a central population register was discussed to serve the needs for a nation-wide system of permanent identification numbers. The Central Bureau of Statistics worked out the plans and was allowed to start the implementation of a central register in 1963.

The permanent identification of persons which was constructed, is a number of eleven digits. The first six digits, called birth data, represent date of birth, with the two first digits of the year omitted. The next three digits are used to distinguish between persons with the same date of birth, between men and women, between people born at the same date in different centuries, etc. The last two digits are check-digits computed according to the modulus eleven method. The number of errors which will pass undetected through the check-computations has been estimated to 1 out of 100 000 errors in identification numbers. This identification number will be used as an internal identifier even though it is not the most ef-

ficient for this purpose.

The establishment of the central registers was done by punching name, address, sex and date of birth for all persons from the census lists. By an automatic processing routine identification numbers were computed and assigned to these cards. The routine keeps an account of which numbers are free and which are occupied for any given date of birth. The numbered cards were distributed to the local registers for control and for acquainting the local offices with the numbers to be used. Births and immigrants are reported currently to the central office which assigns numbers to them. Because of the difference in time between the 1960 Census of population and October 1, 1964, which was the date chosen as a check-point, the control and supplementing work became very extensive. The central register is held on magnetic tape, and one version requires about 30 reels of tape. During the period of establishment and control of the register several hundred reels are in use.

The central register will be kept up-to-date through the regular channels for registering births, marriages, migration and deaths. It will be used for the larger service tasks which have had to be carried out by each local office. From January 1, 1967, the identification numbers will be introduced in tax and social insurance work. As well the police as the health authorities will probably want to use it, and we hope it can be introduced in the school system already at the elementary school level.

During 1968 we plan to extend the central register to a system comprising both a population register and a household register. The necessary data for such an extention are already contained in the local punched card registers.

3.3. Central registers of establishments and enterprises

In connection with the 1953 Census of establishments, a system of central registers of establishments and enterprises were established. Up to 1965, these registers were held on punched cards, but they have now been transferred to magnetic tapes. They comprise most industries, except for agriculture, forestry, hunting and fishery. In some industries the registers do not include enterprises in which the owners are working all alone. In total, the register of enterprises contains about 110 000 units, while the register of establishments covers about 130 000 establishments. The external identifier is usually name and address.

To satisfy the requirement for permanent internal identifiers, independent numbers for enterprises and establishments were introduced in 1965. The identification numbers are assigned continuously with no build in information. They consist of six digits plus one check-digit computed by the modulus eleven method. The check-system will fail to detect 1—3 out of 100 errors in identification numbers. This is assumed to be of no practical significance because these numbers in general will be automatically pre-coded on the documents in most applications and the probability for errors is therefore expected to be small. Integration between units of the two registers is possible

because the register of establishments include the enterprise identification as a link characteristic for each establishment. The introduction of permanent identifiers has initiated a discussion about definitions of "birth", "migration" and "death" for both enterprises and establishments. The two units may have different life-times even though the latter may belong to the former for a certain period. This has been one of the main reasons for introducing independent numbers for an enterprise and its establishments.

The original purpose of the registers of enterprises and establishments was to obtain a tool for performing coordinated statistical surveys, but they have also proved to be of great value for a number of other needs. There are, however, very important administrative needs which cannot be served by these registers in their present form.

A difficult problem in connection with the registers is to get information fast enough about new units. Many sources of information are used of which the social insurance administration is the most important. Maintenance is also based on information collected by mail on "name cards" which are sent to and returned from the units. These cards are very simple questionnaires partly filled out with information from registers by the Central Bureau and which the units are asked to correct and supplement.

In connection with the next Census of agriculture, the incorporation in the registers of units within agriculture, will be considered.

3.4. Register of employers

Tax authorities and social insurance administration need a nationwide register of employers for indirect and direct taxation, collection of insurance premium, etc. The units will partly be personal employers and partly non-personal firms or companies. It would be an advantage if the units may be characterized by the industry to which they belong. The register of enterprises has for this reason been considered as a basis, and it has been proposed that the Central Bureau of Statistics is asked to establish and maintain such a register based on reports from local tax and social insurance authorities.

From statistical point of view this arrangement will be of great interest because it will supply the register of enterprises with units which are impossible to register to-day. The reports will also solve the present problem of registering new units more satisfactory.

The Central Bureau of Statistics has outlined a system including both the register of enterprises, the register of population and a register of employers. It is assumed that the identification number system used for persons should also be used for personal employers and that similar eleven-digits numbers are assigned to companies. This implies that the present identification number for each enterprise will be substituted by the identification number of the person or the company which is the owner of the enterprise. In this way an efficient coordination of identification numbers will be obtained for physical persons, companies, enterprises, employers, tax payers, etc. In other words, each enterprise will get

the same number as the owner, whether he is a person or a company, a taxpayer's identification will be that which is assigned to him in the population registration or, in case the taxpayer is a company, that which is assigned to it in the register of companies and also used in the register of enterprises, etc.

No final decision has yet been taken as to when the work on this project will start and when the system should be operative.

3.5. Central register of land properties

Outside the Central Bureau of Statistics work is going on to construct a register of land properties which will be described by their geographical coordinates and the old system of identification of farms. The bureau has been asked to assist in the technical maintenance of this register.

4. Data collection

Investment in data capital is done by collecting data either *directly* from each individual unit, or *indirectly* for a group of units from an agency which has collected the data for administrative purposes. Data collected for administrative purposes and which are satisfying the requirements of the statistical system will in general be less expensive to copy than to obtain by direct collection.

By a systematic introduction of the central registers and code standards in as many as possible of the administrative processes, we may acquire an increasing amount of useful data without expensive direct collection. In Norway, the register of population will be extensively used in administrative processes. We hope that the same will happen when we manage to build a register of employers into the register system. We are spending great efforts to establish central registers and to give service from them. We hope it will result in an increasing stream of useful and well identified data to the Bureau.

Some of the data collection done directly to-day will then probably be superfluous. Already now, we can assume that the censuses don't need to cover the same width of questions as previously. The censuses will probably more be aimed towards control of the coverage of the registers and to get data which cannot be registered currently.

The central registers will therefore limit the questions to characteristics not yet registered. In a Census of fishermen which is carried out during the winter of 1966/67 only data which are not known from the 1960 Census of fishery or the current population registration are collected. Those data collected will for each person be linked with data from the 1960 Census, etc., and this combined data set represents what would have been necessary to collect without the permanent identification system.

5. Data files

5.1. Filing order

The concrete counterpart to the theoretical concept data capital is *data files*. There are many alternatives for

ordering the files, but minimum conditions of the statistical file system are that any observation value must be accompanied by permanent unit identification, standard code for the characteristic and time specification. The files may be ordered by *unit, subject* or *time* depending on which of the three types of descriptors was used as main sorting key. The optimum order depends on both the order of the data collected and the order required by the computation of statistics.

In the Central Bureau of Statistics the data files are kept on magnetic tapes which so far seem to be the most appropriate medium for this purpose. The data are kept on tapes ordered by time/subject with unit ordered data within each tape.

5.2. Files of population data

The central register of population gives us demographic data for each person from 1960. Relations between parents and children are also filed. Such relations permit the description of units by a number of indirect characteristics. Children may, e.g., be characterized by the structure of their families. The registration of births and deaths also includes a number of medical data. The data received will give new and more extensive possibilities for testing and analyzing theories for births, marriages and deaths. In the same way, the migration may be studied more intensively than what has been possible so far. There should therefore be no need in 1970 to carry out a Census of population as comprehensive as the previous censuses.

From 1967 tax and income data for all individual taxpayers will be identified by the central population register number and can be linked with data from files of population. This will give a new basis for analyses of income and wealth and the possibility to study the effects of political actions on the units. It will also be possible to study the changes in the frequencies of marriages and births in relation to the economic conditions of the individual families.

Later on we shall be able to include data on social, health and educational background into the system. We shall in this way obtain a better basis for sociological research as well as for socio-economic studies of the interaction between sociological and economic factors. Among concrete projects discussed, analysis of death probabilities by profession and life pattern, and models for choice of education by income and social status of parents may be mentioned.

If the central register of employers is established as outlined above, information reported about taxes which the employers must deduct from each individual salary and wage, will give a data set connecting any employee to his employer. These data will make it possible to use the conditions within the enterprise as explanatory factors for the behaviour of employees in sociological studies.

The central register of land properties and the register of population contain both the old farm identification number and a description of location or home address.

The first register will in addition

have geographical coordinates and we shall therefore have the possibility to characterize a large part of the population by the approximate coordinates of their homes. In statistics based on persons, we may therefore define geographic areas, independent of the present administrative areas, which are assumed to be of particular value for geographers and those developing plans for smaller regions.

Administrative sources will to a large extent supply the files with data for persons. One characteristic which so far seems to be difficult to register currently is, however, profession.

5.3. Files for enterprise and establishment data

The introduction of permanent identification numbers for enterprises and establishments permits the organization of files for enterprise and establishment data from 1964 with basis in the data from 1963 Census of establishments. For a group of 600 enterprises comprising 1 300 large establishments in 1963, we are now establishing special individual time series back to 1959.

These data files give us already to-day a readiness for describing the number and distribution of units which is regularly published. When the individual time series in the files grow longer, they will be utilized for analyzing the "births" and "deaths", the life-cycles, etc. of enterprises and establishments. Studies already being done, indicate that the age-distribution of establishments may, e.g., be a much more interesting classification than what has been realized so far. Another field of great interest is the computation of product functions including time-lags, studies of investment and other behaviour functions connected to the enterprises and establishments.

A large amount of data for the establishments will also in the future have to be collected directly by the Central Bureau of Statistics because they are not required by any administrative process. But if the register of employers and the register of enterprises become coordinated in one system, important identified data about enterprises can be obtained from administrative processes, mainly from tax authorities. In addition to data for personal tax payers, identified income and tax data for companies will be available for linking to other types of enterprise data. Also accounting data for enterprises in industries for which such data are required by tax authorities, may be available. They may probably become a very useful supplement to or perhaps a substitute for directly collected data on company accounts. A coordinated system of registers of enterprises and employers will make it possible to link data on the indirect taxes currently collected to other data for enterprises and a better basis for short-term statistics on retail sales may be obtained.

The data for tax deductions in salaries and wages mentioned above may also be used for linking the personal characteristics of enterprise managers to their enterprises. We may then study the hypothesis that the managers' education and experience affect the development of their enterprises.

In general, the data files will prob-

ably lead to a greater interest for including in models and analyses the effect of different distributions.

6. Computation of statistics

The utilization of data files and statistical files for deriving aggregates of data for groups of individual units is here called *computation of statistics*.

Computation of statistics has frequently been organized on the assumption that data could only be utilized efficiently once and therefore the computation process should result in statistics which were as general purpose as possible. Analysts have mainly made their computations on aggregates. The data files will allow repeated computations on individual data and the needs for computing a large number of general purpose statistics will decrease. Instead, efforts can be concentrated to more intensive computations to satisfy specified analytical tasks. The data files will also give a much more extensive material for precise estimation of coefficients and permit that relations which cannot be estimated from aggregates, will be included in the models.

In Norway we have little to report yet about experience in utilizing data files, but we are at the moment discussing several new projects based on the new possibilities represented by the data files.

7. Statistical files

The capital of statistics is maintained in *statistical files* in which statistics are identified by permanent identifiers for groups, characteristics and points or periods of time. Such statistical files are often called statistical banks and are kept on punched cards and magnetic tapes in order to obtain fast access to requested statistics.

Foreign trade statistics and National accounts are the fields in which we have most experience with statistical files. For Foreign trade statistics, a file with statistics for more detailed groups than those which are printed, is updated every month. This file is utilized in a subscription system. National accounts were kept on punched cards for many years and are now transferred to tape which gives an efficient access for further computation based on National accounts.

The statistical files may technically be kept for fast access in a data processing system with inquiry stations connected from which any filed statistical data can be retrieved in seconds. This may have great implications for the future statistical information service.

8. Final remarks

We have discussed some problems which arise in connection with the implementation of those ideas which have been called a statistical file system. There are a number of problems which have been left out because of limited space and time. Two further problems must, however, be mentioned.

The great importance of central and permanent unit identifications required by the statistical file system has been emphasized. Equally important is the standardization of definition of characteristics in such a way that data from

different sources will be consistent with each other, just as the concepts of the National accounts have been standardized and made consistent. A part of this task has been solved through the construction of statistical standards of classification, but there are still much to be done in order to get these standards introduced in administrative processes and there are also many fields for which standards have not yet been developed.

The second problem to be mentioned is the risk that the files may be misused. A positive reaction from the public on the statistical information system with its files will probably depend to a large extent on whether the system is developed in such a way that the risk for misuse is reduced to an unsignificant level.

9. References

(1) E. S. Dunn, Jr.: Review of Proposal for a National Data Center, Bureau of the Budget, Washington D.C., 1965.

(2) S. Nordbotten: A Statistical File System, Statistisk tidskrift, No 2, Stockholm, 1966, pp. 99—103.

(3) Congrès International de Statistique: Registres de Population, Rapport de IIme Section, Huitième Session, St.-Pétersbourg, 1872.

KUNGL. BOKTR. STHLM 1967