Arbeidsnotater

STATISTISK SENTRALBYRÅ

WORKING PAPERS FROM THE CENTRAL BUREAU OF STATISTICS OF NORWAY

IR 67/2 Oslo, 22 May 1967

PURPOSES, PROBLEMS AND IDEAS RELATED TO STATISTICAL FILE SYSTEMS

BY Svein Nordbotten

Central Bureau of Statistics, Norway

Content:

Invited paper to the 36th Session of the International Statistical Institute in Sydney, 28 August - 8 September, 1967.

1. Introduction[1)]

By statistical file systems we mean systems based on systematic preservation of acquired data with the main purpose of responding efficiently to the continously increasing demand for knowledge about the development and state of being of the society in general. The components of such systems comprise data acquisition, data storage, data processing and information supply. Important tools for modern statistical information systems are the automatic data processing systems which have made it possible to store large masses of knowledge in readily accessible files and link related data to an extent which we only a few years ago were unable to imagine.

Many of the ideas of the modern statistical information system were, however, conceived years ago. In the 8th session of the International Statistical Institute in St. Petersburg in 1872 a special section was devoted to the statistical possibilities of population registers. The views presented nearly a century ago have a great similarity to those we present to-day in our discussion of population registers. The great difference, however, is that to-day we may be able to implement these ideas.

The needs for more efficient statistical file systems have been discussed for some time by users of statistics. The Social Science Research Council in the United States established the Ruggles Committee to study the preservation and use of economic data and the committee presented its report in 1965.

The International Social Science Council has sponsored two International Conferences on Data Archives in 1963 and 1964, respectively. The problems were discussed at the Nordic Statistical Meetings in 1960 and 1964 as well as in many other connections and places [4,5].

These discussions have been followed by studies of means for preparing the statistical systems to meet the requirements of the future. In the United States, the National Bureau of the Budget at the end of 1965 presented the Dunn report reviewing the proposal for a national data centre [2]. A special task force committee, the Kaysen Committee, discussed and recommended the establishment of a federal statistical system for the US in 1966. In Sweden a special committee is working with the problem on a broad basis. Many national, statistical offices are approaching the problems. In 1967,

1) The aim of this paper is to present ideas, concepts and principles from the Scandinavian discussions on statistical file systems. Examples of practical application and experience are discussed by Mr. Ingvar Ohlsson in his paper [7].

related problems have been discussed in a technical group established by the Conference of European Statisticians, at a conference in Rome organized by the International Federation for Information Processing and the Federation Internationale de la Documentation, at a conference in Jerusalem organized by the Information Processing Association of Israel, etc.

2. Theoretical considerations

2.1 Concepts

A file system is composed of three main components, i.e. sets of data, procedures for acquiring and processing the data, and technical processing devices. In this paper we shall discuss the two first components.

Data are standardized representation of recorded facts. Data must comprise references to the objects to which the facts are connected, descriptions of the facts in question, and time specifications. The object or set of objects to which data refer, is called a unit. The unit must be defined in such a manner that it can be recognized and identified. It may be a person, a professional group of persons, an establishment, an industry group of establishments, etc. Which types of objects are recognized as units in a statistical system vary by the structure of the society served by the system, the costs of acquiring, storing and processing data for the different types of units and by the demand for statistical information.

The description of the fact may be divided into the characteristic which describes the kind of fact recorded, and the value which specifies the description. The last component of data is the time specifications which refers basically to a point of time or the intermediate period between two points of time.

Data may be denoted in a symbolic manner by the vector

$$d = (i, k, t, x)$$

where i represents the unit identification, k the characteristic, t the time specification and x the value of the characteristic.

We may distinguish between observed data based on a direct observation or recording of facts associated with a unit and computed data which represent facts obtained by computations on other data from the same or related units.

Data capital is the preserved stock of collected and computed data and plays a central role in a statistical file system similar to that of the production capital of an industry. Its productivity depends on the degree to which data are organized in a manner which satisfies the requirements of

the systems. These requirements may be illustrated by a data-box containing a number of small rooms for storing values. Each data item is identified by the statistical unit, i, to which it is associated and which has its permanent position along the first axis of the box, by the characteristic, k, observed or computed, which has its permanent position along the second axis and finally by the period or point of time, t, which has its position along the third axis of the box. The value, x, belonging to the data is stored in the room determined by the coordinates i, k,and t (Figure 1).

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. We may express this symbolically by

$$S_1 = t'((i_1,k_1,x_1) \ldots\ldots (i_n,k_n,x_n))$$

where t' is the common time specification and n is the number of data stored for t'.

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 which may be denoted by

$$S_2 = k'((i_1, t_1, x_1) \ldots\ldots (i_{n'}, t_{n'}, x_{n'}))$$

and

$$S_3 = i'((k_1, t_1, x_1) \ldots\ldots (k_{n''}, t_{n''}, x_{n''}))$$

The data-box organization requires therefore that we have a system of permanent unit identifiers, standard codes for all characteristics and time specifications.

Data supplied by means of the data capital to satisfy a need for knowledge are referred to as information. By statistical information is meant data which are considered to describe the state of being and the development of the society. This implies that the data capital may very well contain data which cannot be supplied as statistical information. The extent and delimitation of statistical information in contrast to other kinds of information, may vary with the technical development and complexity of the society, the legislation as well as the particular traditions of the respective society.

By investment in data capital we indicate the processes which are needed to increase the stock of data. There are two ways of doing this, either by data acquisition or by data computation which results in observed and computed data, respectively. The data acquired may either be collected

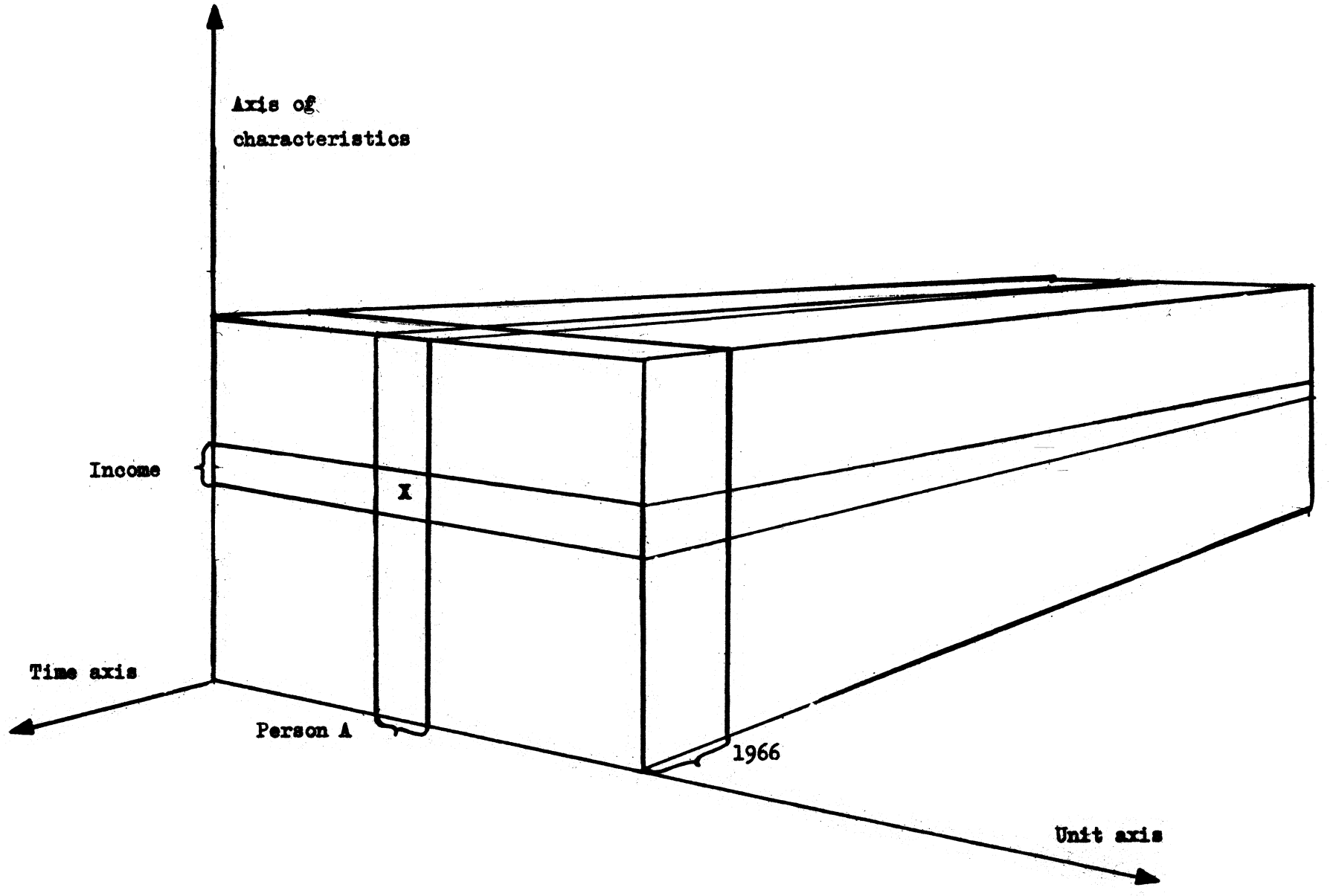

Figure 1 : The data box and the storage of income X for person A in 1966

for statistical purposes exclusively, or be transferred from agencies collecting data for administrative purposes.

Supply of information is the information delivered from the data capital. It may either be performed by retrieval and presentation of data as a response to special purpose demand, or by retrieval, presentation and multiplication for satisfying general purpose information.

2.2 General model

The main characteristics of a statistical file system may be summarized by the following relations. The statistical information supply, G_t, during period t is a function of the accessible data capital, D_{t-1}, at the end of period t-1, and a utilization factor, u_t, indicating the degree of publication of information. This relation is denoted by:

$$(1) \qquad G_t = G(u_t, D_{t-1})$$

Register service, R_t, supplied from the statistical file system to agencies collecting data for administrative purposes, is also assumed to depend on the size of the data capital at the end of the previous period and a service factor, r_t, expressing the effort utilized in applying the data capital for this purpose. Let this relation be:

$$(2) \qquad R_t = R(r_t, D_{t-1})$$

The size of the data capital, D_t, at the end of period t depends on the size at the end of the previous period, D_{t-1}, the investment, E_t, in computed data and, C_t, in acquired data, between the two points of time and a data organization factor, d_t, representing the efforts made in organizing the data in a productive data capital. This relationship is represented by:

$$(3) \qquad D_t = D(d_t, C_t, E_t, D_{t-1})$$

The computed data, E_t, are obtained by utilizing the data capital at the end of the previous period for deriving new data by computation. The computation activities are denoted by e_t, and the relationship by:

$$(4) \qquad E_t = E(e_t, D_{t-1})$$

The acquired data, C_t, are dependent on the register service supplied in the previous period resulting in a transfer of acceptable data from administrative agencies, and on the statistical collection, c_t. The relationship

is:

(5) $C_t = C(c_t, R_{t-1})$

The cost, K_t, of the statistical system in period t will be assumed to depend on the cost of storing the data capital in an accessible way, the statistical collection and the computation activities, the organization efforts, the register service and the information utilization:

(6) $K_t = K(u_t, r_t, d_t, e_t, c_t, D_{t-1})$

One objective of the statistical system may be to develop a policy for period 1 up to T expressed by values of u, r, d, e, and c for each period given an initial D_0. The policy preferred may be that which maximizes a function:

$$W = W(G_1 \ldots\ldots G_T, K_1 \ldots\ldots\ldots\ldots K_T)$$

expressing the evaluation of gains obtained by the society by means of the information supplied and the costs expended on the statistical system.

The problem of developing the preferred policy will be a problem of dynamic optimization which will require the numerical spesification of the relations. We may, however, make some general observations:

1. The data capital is a central factor in the system. The size of it will according to equation (3) depend on how well it may be preserved from depreciation. This is in this context partly a problem of efficient data organization which is discussed in section 3 below.

2. Data are partly acquired from administrative agencies and partly by direct collection. According to equation (2) and (5) the former may be increased by extended register service. These problems are discussed in section 4.

3. The data capital may be increased by storing collected and computed data and utilized by copying data stored. The possibilities for investment by computation is according to (4) dependent on the size of the capital. These processes are discussed in section 5.

4. The information supply represented in (1) is finally discussed in section 6.

3. Data file

3.1 File structure

The data file is the concrete counterpart to the theoretical data capital. The basic component of the file is the record which may be considered as composed of two logical fields. The entry field identifies the record within the file. The content of the entry fields may e.g. correspond to the coordinates (i, k, t). It may be omitted if the record is identified by position. The second field is the exit field containing the information corresponding to the value x. The records may be organized in sets corresponding to a slice or a linear subclass of the data-box, e.g. a set may represent all records with data referring to the period t'. In this case the entry fields of the records need only to comprise (i, k). A set may be a record or may comprise the whole file and is related to the particular application.

The data file may consist of a hierarchy of data sets. One set with all existing time specifications, t, in the entry fields, may for example be established. The exit fields contain links, l, to the data sets corresponding to the respective time specifications. The entry fields of these sets need therefore only to comprise (i, k). We may term t the symbolic name and l the link to the data set referring to the period t. The link l may also be called the label of the data set with the symbolic name t.

The data file structure may be illustrated by a file structure tree with a root symbolizing the initial set and branches which end with the terminal sets containing the desired data. This implies that a data set is identified stepwise through a chain of links within the data file starting from the initial set. The set of time specifications may be regarded as an initial set. A three level structure is for example obtained if for each time specification a second level set is established with the entry fields which contain k and the exit fields which contain the links to data sets with entry and exit fields containing i and x, respectively (Figure 2). The purpose of the construction of such a file structure is to reduce the search necessary to retrieve a desired set.

3.2 Registers

The registers of a statistical file system are the means for moving to the correct position along the axis of units. They may be considered as a data set in which each record contains identification data for a unit at a certain point of time. The register is organized in the file through the

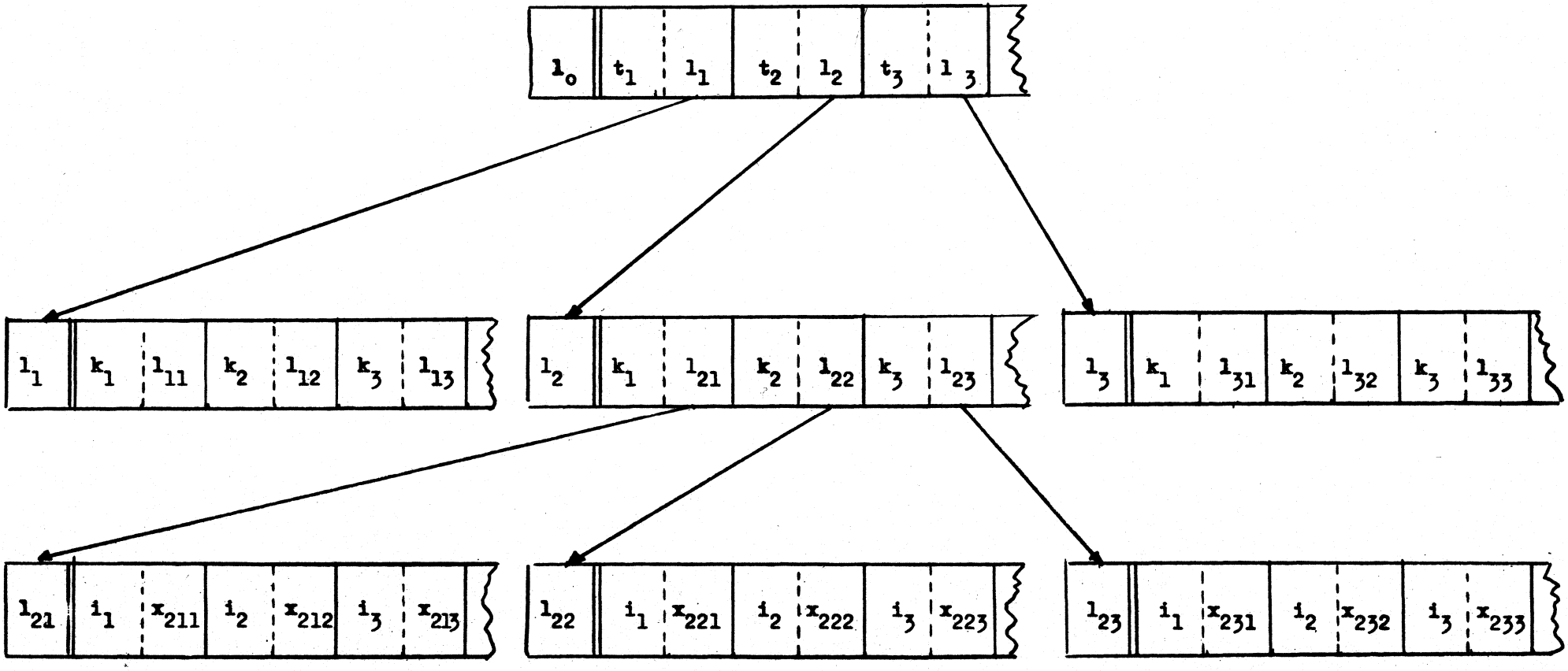

Figure 2 . File hierarchy

first level symbolic name, t, specifying the point of time relevant and the second level symbolic name k_r which specifies that the characteristic in question is the name and address of units. The symbolic name of the register is therefore (t, k_r). The content of the records of the register is i in the entry fields and x in the exit fields. In this set the value, x, represents the information needed to identify the unit in the real world. We may denote this the external names of the units. The permanent unit identifier, i, should ideally carry no information in order to avoid occupying unnecessary storage space in all entries in which it occurs and in order to keep the internal identifier permanent in spite of changes or errors detected in the recorded characteristics. The usual practice of including the date of birth in the population identification numbers does not always seem to be appropriate since we may find that the reported date of birth was incorrect.

The maintenance of such registers may raise quite new definitional problems related to concepts of birth, migration, marriage, divorce, death, etc., for units such as establishments, enterprises, and municipalities. The external name of the units may change and this aspect must therefore be carefully observed. This may require continous and expensive maintenance efforts which may be difficult to motivate on the basis of statistical needs only. A register is therefore formally up to date only for a specific date. Applications of a register as a means of making connections between the real world and the file unit for any other date will therefore imply an additional risk for incorrect linking [1].

3.3 Description sets

A set with entries containing the characteristics, k, may be called a description set because its records give general descriptions of the characteristics to which they refer.

Frequently a data set may be of the compound type denoted by

$$S = ((i_1,k_1,\ldots,k_K,t,x_{11}\ldots x_{1K}) \ldots\ldots (i_N,k_1\ldots k_K,t,x_{n1}\ldots x_{NK}))$$

which means that K characteristics $k_1 \ldots k_K$ are recorded for eaeh of the N units $i_1 \ldots\ldots i_N$. If all records have the same time specification, t', such a set may be identified in a file hierarchy by the symbolic name $((t',(k_1,\ldots\ldots,k_K))$ through the initial time reference set which for entry (t') refers to a description set in which the entry $(k_1,\ldots,k_K)$ refers to the third level terminal set with records containing $(i_1, x_{11},\ldots x_{1K}) \ldots$ $(i_N, x_{N1},\ldots\ldots,x_{NK})$. This we may denote by

$$S = t'(k_1,\ldots,k_K((i_1, x_{11},\ldots,x_{1K}) \ldots (i_N, x_{N1},\ldots,x_{NK})).$$

Let us suppose that there are three characteristics for each unit, e.g. k_1 = sex, k_2 = profession, and k_3 = income. We may want to organize the file in such a way that we may identify the set consisting of females within a certain profession only. This may be obtained in the following way.

In the second level set referred to by t', one record with entry (k_1, x_1) for each of the alternative values of x_1, and similarly one record with entry type (k_2, x_2) for each of the value alternatives of x_2 are introduced. If we have 100 professions we have (100 professions) + (2 sexes) = 102 entries in the second level set. For each of the 100 records with entry type (k_2, x_2), one new third level set is introduced containing only two records with entry field type (k_1, x_1) and for each of the two second level set records with entry type (k_1, x_1), one new third level set with one hundred records with entry field type (k_2, x_2) is introduced. These, in total 400 third level entries, will refer to 200 fourth level terminal sets all with record type (i, k_3, x_3) each representing one of the 200 possible combinations between sex and profession. This organization identifies all logical sets defined by sex and profession.

A very important aspect is how to construct the descriptions of characteristics in order to be able to retrieve the data from the file. We shall return to this problem when discussing filing processes below.

3.4 Security sets

A very much important question is how to safeguard the public against misuse of the data of the file. It should at once be emphazised that in any statistical system there will be a risk that it may be used in a way which may injure individual interests or rights of privacy. To keep this risk below an acceptable level, laws and regulations have to be instituted defining which information should be considered confidential and therefore never supplied from the statistical system, which may be supplied to one special class of users but not to others, etc.

A security set is a set by which some of the intentions of the regulating laws may be incorporated in the file system. The entries of a security set contain secret passwords known only to authorized users. The password may be considered to be a special characteristic, p, associated to the data of a set. If the value, z, of this password is required to identify the set it is said to be locked.

As an example, consider a data set with records of the type:

$$S = (t, (k, (p, z'))) \cdot ((i_1, x_1), \ldots\ldots\ldots, (i_N, x_N))$$

where z' denotes the secret value of the password while i and x are the unit identifier and the value, respectively, for the N units recorded. If this data set is organized in the file through an initial first level time specification set, a second level description set, a third level security set which finally refers to the data set containing records with (i, x), this data set is locked and unavailable without the knowledge of the correct entry (z') of the third level security set and which is a required part of the symbolic name of the wanted terminal set.

The idea of locked data sets may of cource be extended to multiple level security sets requiring the consent of several authorized persons for use. In statistical systems of a decentralized type in which several data collecting and processing agencies share a common file, the security sets may be an important part. In centralized statistical systems it may on the other hand be less relevant.

3.5 Terminal sets

The terminal sets will in a multi-level file usually contain records of the type

$$(i, x_1, \ldots\ldots\ldots, x_K)$$

The way data are kept in the file may also give rise to important distinctions. There are some set which may be called the status sets which contain data describing units at a certain point or period of time, i.e. all data refer to the same time. This does not imply that they only refer to what is often called stocks. The other class of sets are the change sets which contain data about time specified changes in the characteristics of units during a limited period. In a change set the time specification within the period may vary among data. While a status set gives a description of all units within a group and the same characteristics for each unit, the change set only contains data on those units and characteristics for which a change has been recorded. A population register per a certain date is a typical status set, while a set containing the births, deaths and other changes during a period is on the other hand a change set. There should always be consistency between two status sets and the corresponding change set for the intermediate period.

In some situations it may be appropriate to consider whether absolute or increment values should be stored. In a system in which absolute values are stored, increment value sets can be obtained by taking differences between absolute values. On the other hand, a system with increment value sets will always require at least one absolute value set and by cumulation other absolute value sets can be obtained.

3.6 Optimum file organization

The file organization has several aspects which need to be considered in the search for the optimum file organization. These problems will, however, be related very much to the local conditions including the file equipment available and are therefore only surveyed very briefly here.

The first problem is how the file ought to be partitioned into physical data sets in order to draw maximum efficiency out of available equipment subject to the composition of the file work. A second problem is how to design the file hierarchy structure. A third problem is the arrangement of the records within a data set. A fourth problem may be to which extent identical data should be dublicated or allowed to be present in different physical data sets to make them as compatible as possible with logical data sets.

The objectives of the file organization may be stated in different ways. One possible objective may be a file organization by which the expected, average storage and retrieval cost or time is minimized subject to given retrieval efficiency, specifications and available equipment. This organization will of course partly depend on the composition and frequency of demand for the different data set. A very challenging task in this connection would be the construction of a self-optimizing file system which kept track of the use of different sets and adjusted the organization according to the frequency of use [6].

4. Data acquisition

4.1 Statistical collection and transfer of administrative data

Data acquisition represents all kinds of activities which bring new data into the statistical system and represents the first operation in what was called data investment.

We may distinguish between statistical collection and transfer of administrative data. By statistical collection we have in mind data collection

carried out at the expense of the statistical system for the direct purpose of investing in the data capital. Population censuses, annual surveys of manufacturing, etc., are among the well known and usual examples of direct collection for statistical purposes. Transfer of administrative data is the flow to the statistical system of data collected for non-statistical purposes by administrative authorities. Use of income data as obtained from tax authorities in some countries is a typical example of this type of data.

From a statistical as well as from a national point of view, acquiring administrative data is less expensive than direct statistical collection of equivalent data. Statistical collection means in many cases a collection of data in parallel to similar data also collected for administrative purposes which entails collection expenses for the statistical system and unnecessary reporting loads on the public.

The condition for the use of data collected for non-statistical purposes is, however, that the data are in accordance with the concepts of the statistical system. In particular, the units referred to must correspond to the units of the file system and without too high a cost be possible to match with these. The characteristics registered must satisfy the needs of the statistical system and be compatible with the characteristics registered by other data collecting sources. Particularly, it seems to be important to aim at a situation in which the same central registers are used both by the statistical information system and the administrative agencies. Such a policy will reduce the matching problem and increase the efficiency of the registers.

Using the registers in data collecting processes, it is important to pre-identify the questionnaires or forms to make the identification of data as efficient as possible. When pre-identification is impractical, self-identification would be desirable. Both approaches assume that an identification number with a one-to-one correspondence to the internal identifier is used. Self-identification requires of course that the data source knows its own identification number. Pre-identification may be used in annual surveys of industries when the questionnaires are mailed out by the use of a register. In addition to the name and address of the establishment, also its permanent identifier, i, must then be printed and used when the form is returned. Self-identification may be used in administrative processes. When applying for admittance to a school, it may for example be required that the applicant gives his identification number on the application form.

Errors may occur during transfer of the identification numbers, e.g. by punching of data into cards. To reduce that risk of erroneous identification

because of such errors, identification numbers, i', often include one or several check-digits, i'', for checking the validity of the numbers possible. A check-digit may be regarded as a functional value, $i'' = f(i)$, of the original identifier, i, and contains no additional information. The form of such identification numbers is therefore $i' = (i, i'')$. The computation of the check-digit may be repeated as a check of the validity of the number during or after a transcription process. Check-digits may only catch special types of errors and can by no means eliminate the problems of erroneous identification numbers.

Subject to a certain error structure, self-correcting identifiers may be constructed based on additional digits by which it is possible to correct errors appearing in the numbers.

The check-digits are frequently used together with the identification number also within the file, i.e. the identifier i' is in fact also used as internal identifier. With the present level of precision in the work of automatic data processing equipment there is no need for such means for checking the storing and retrieval procedures. The check-digits should therefore be regarded as an annex to the data which need not to be stored since they do not represent any additional information.

The process of identifying data may be considered as a problem of matching data against a register, the solution of which may be to construct matching procedures minimizing the risk for mis-matching. The more information available for each unit in the register the better the chances for correct matching may be. This and related problems are discussed by Fellegi and Sunter [3] who present a very interesting approach to the solution of the problem of optimal matching. The matching problem may seem to be eliminated where pre-identification is applicable. But the truth is usually that the matching problem is then loaded on the mailman, or the interviewer. Even when self-identification applies, the risk of mis-matching is not eliminated. From identification point of view the registers should therefore comprise as much data about the individual units as possible. The registers will, however, grow clumsy very quickly. A more satisfactory approach may therefore be to dump the data into the file and for each application retrieve just that data which are needed for matching and identification.

4.2 Continous and non-continous collection

Collection of data may either be continous or non-continous. In a continous collection scheme events are registered, and the data collected

when they occur, while in non-continous collection all units are observed and their characteristics registered at certain points of time. Continous collection will usually result in cumulating change sets and non-continous collection in status sets.

The non-continous collection scheme has the advantage that for a description of the situation at enumeration time, t', only, there is no need to keep track of the individual units between two collections, i.e. the non-continous collection scheme does not require the permanent identification system for satisfying the simple needs. If, however, analytical requirements assume that the individual data of two collection processes can be related, the permanent identification system will be a necessity. Still, there will be no need for recording the time for each unit and usually the characteristics recorded are the same for all units. We may therefore organize the data in for example the form

$$S_1 = (t' (k_1,.....,k_K))((i_1, x_{11},.....,x_{1K}).....(i_N , x_{N1},...,x_{NK}))$$

where the symbolic name $(t'(k_1,....,k_K))$ represents the terminal set.

The advantages of a continous collection scheme are that only changes or new events have to be recorded and that the situation at any point of time may be described. A data set obtained by continous collection during a period will, however, require time and characteristic specification for each record which normally only will contain a change in one characteristic:

$$S_2 = ((i_1, k_1, t_1, x_1) ..... (i_M, k_M, t_M, x_M)$$

A statistical file system is based on the idea of permanent identifications, and the continous collection scheme will probably play a relatively important part of such a system. This is supported by the fact that most administrative needs are focused on the recording of changes.

The continous collection scheme contributes to a more balanced distribution of collection activities over time and to the reduction of the number and extent of large censuses. There are, however, characteristics particularly those for which there are no contemporary administrative needs, which may be most efficiently collected in a census. The censuses may also provide convenient status sets for checking the current registration.

The statistical file system may also create some interesting problems of sampling since the objectives as well as the basis for sample surveys may be different from those which have so far received most attention.

4.3 Optimum data collection programmes

The above discussion indicates that the optimum data collection in a statistical file system will represent new and challenging theoretical and methodological problems to which we so far have paid very little attention.

5. Data processing

5.1 Processes

In addition to data acquisition a number of other processes are performed in connection with the investment in data capital and supply of information. We shall discuss three processes, the storage, the retrieval and the computation of data sets, which are the main aspects of the processing in a file system.

Each of these processes comprises standard processing operations such as sorting, merging, matching, extraction, reproduction, etc. of data sets. A discussion of these operations is outside the scope of the present paper and is not included here.

5.2 Storage

Data storage is the process by which collected or computed data sets are preserved for the future. The data storage starts by the standardization of the description of facts collected or computed. The standardization is also called conversion, coding or classification. It is performed in order to make the processing more convenient and in order to be able to recover the data when needed. The price to be paid for this is a loss of details, and it is important to find a good balance between gains and losses.

The result of the standardization is data of the form d = (i,k,t,x). As pointed out above data will frequently contain compound characteristics because several characteristics are usually recorded simultaneously. The characteristic k may be rather complex taking care of different kinds of characteristics and relations. The fact, F = (establishment (A) in a period (T) sold (S) amount (X) measured in pieces (P) of commodity (C) to (R) establishment (B)), may seem quite simple, but exchanging the word "sold" by "purchased", or "pieces" by "dollars", etc. will change the meaning completely. We may solve this by considering k and x as vectors whose element has the following meaning: k_1 = action, x_1 = S, k_2 = quantity, x_2 = X, k_3 = measurement unit, x_3 = P, k_4 = commodity, x_4 = C, k_5 = relation, x_5 = R, k_6 = unit, x_6 = B. We specify i = A and t = T. If we in addition introduce

some rules about the sequence, we may by a record of the form (i, k_1, ..., k_6, t, x_1,.....,x_6) have saved the data of the fact described with the accuracy permitted by the feasible alternative values of the x's.

The means at our disposal for standardization are registers with permanent identifiers and systems of standard classifications. Together with the rules for how these are used in standardizing fact descriptions as data they may be called a data description language in which the registers and systems of classifications represent the vocabulary and the rules are the syntax of the language. Particularly the systems of standard classifications need further development and consolidation to satisfy the requirements of the file system.

The second step of storage is the inclusion of the data set in the file. Depending on the structure of the file, whether the data set is locked, etc., special routines are needed for updating the sets at different levels with new records whose entries are elements of the symbolic name of the data set, for generating links to the next level set in the exit fields of the new records and, finally, for including the terminal set.

Consider as an example a three level file with an initial time specification set, a second level of description sets and a third level of terminal sets. Let

$$S = (t' (k'))((i_1, x_1),....., (i_N, x_N)$$

be a set to be stored. First, the initial set must be searched for t'. If this entry already exist, it gives a reference to a description set. In this description set a new record with the entry k' and an exit with a generated identification of the terminal set is established. In case t' is not matched with any entry, a record with the entry t' is established and a link generated in its exit field. This link is the identification of a new second level set which so far only will contain one record whose entry and exit will be k' and a generated link which becomes the identification of the terminal set. Finally, the terminal set $((i_1, x_1)$ ..... $(i_N, x_N))$ is given the generated link as identification and included in the file.

The procedure will, of course, increase in complexity with the levels of the file structure, etc.

5.3 Retrieval

The process which may be considered the reverse to the storage is the retrieval process. Much work has been done in developing theory and systems

for what is usually called information retrieval. This work has particularly been associated with the problems of retrieval of written texts, documents, books, etc., in libraries. To emphasize that our problem is a slightly different data retrieval is used here to indicate that we may want the smallest data element retrieval not only for supplying information, but also for the basis of further systems processes.

Data retrieval is a process in which a description of a hypothetical data set is specified in order to retrieve the data stored, if any, corresponding to the description. The first problem is obviously to specify a description of the wanted data set in such a manner that the retrieved data set will contain, as much as possible, relevant data without exceeding a predescribed limit for inclusion of irrelevant data or vice versa. The percentage of relevant data obtained is an indicator of the retrieval efficiency. The answer to this specification problem may be the data language approach and the problems are much the same as already discussed.

Consider again as an example a three level file. Suppose that we want to retrieve data concerning the characteristic income, k_1, related to the period t' for persons which the characteristic age k_2 having the value x_2'. The name of this hypothetical data set may be written:

$$(t'(k_1, k_2, x_2'))$$

Let us suppose that in the file we have , however, only the following two sets: $(t' (k_1))$ and $(t' (k_2))$. To retrieve the wanted set we shall therefore need a procedure of the following type:

First, the initial set is searched for t' which results in a link to a description set. This is first searched for the entry (k_1, k_2, x_2') without result. Then it is searched for the entries (k_1) and (k_2, x_2') with a link to data set for (k_1). The search is continued for the entry (k_2) with a reference to the second data set. The specification is not yet exhausted and the set with the name $(t' (k_2))$ which contains records of the form (i, x_2) are searched for the entries with $x_2 = x_2'$ giving a set of identifiers, i, as links to records within the stored data set with the name $(t (k_1))$. These records form the desired data set and are retrieved.

Keeping data sets stored generates storage cost. Sometimes it may be necessary to delete obsolete data sets. This will require procedures similar to storage and retrieval procedures.

5.4 Computations

The computations comprise a number of functions and applications of data from the data file, e.g. transformations, editing, linking, aggregation, estimation, prediction, etc. Many of the computations will be affected by increased possibilities obtained by the organization of data. We shall only try to emphasize the main consequences for computations because of the file system.

Consider a class of N units, e.g. persons whose behavior we wish to study. Our hypothesis may imply a relation $x_{i1} = f(x_{i2})$ between the values of characteristics, k_1 and k_2, which may as an illustration be the events of being married or single, and the value of income. Let

$$S_1 = t'((i_1,k_1,k_3,x_{11},x_{13}) \ldots\ldots\ldots (i_N,k_1,k_3,x_{N1},x_{N3}))$$

and

$$S_2 = t'((i_1,k_2,k_3,x_{12},x_{13}) \ldots\ldots\ldots (i_N,k_2,k_3,x_{N2},x_{N3}))$$

be two separately collected data sets both with time specification t'. By means of the permanent identifiers, we may retrieve a third set by record linkage,

$$S_3 = (k_1,k_2,t')((x_{11},x_{12}) \ldots\ldots\ldots (x_{N1},x_{N2}))$$

which is what we need for our tests and estimation of the relation $x_{i1} = f(x_{i2})$. A number of practical applications of record linkage is described by Mr. Ohlsson [7]. Without a file system with permanent identifiers, we have to look for a common characteristic, k_3, say age, recorded in both sets. If the values of k_3 are different for any two units, they are in this connection equivalent to the identifiers. This is, however, usually an exception and is not usually true for age. Let us assume that k_3 has $M < N$ different values. Still under certatin conditions a satisfactory record linkage may be possible [3], but often we have to form M new units each defined as the class of all persons characterized by a given value of k_3, i.e. an age group, and we obtain the data set:

$$S_4 = (k_1,k_2,k_3,t')((x_3^1)(x_{11},\ldots\ldots,x_{N_1 1},x_{12},\ldots x_{N_1 2})\ldots\ldots \text{(cont.)}$$

$$(x_3^M) \cdot (x_{11},\ldots x_{N_M 1},x_{12},\ldots,x_{N_M 2}))$$

in which we are not able to make pairs of the x_{i1} and x_{i2} values. We may, however, compute the aggregates. Let $\bar{x}_{j1}$ and $\bar{x}_{j2}$ denote the frequency of persons which were being married and average income, both in age group j during period t'. We then obtain the set:

$$S_5 = (\bar{k}_1, \bar{k}_2, t')(\bar{x}_{11}, \bar{x}_{12}) \ldots\ldots\ldots (\bar{x}_{M1}, \bar{x}_{M2}))$$

where $\bar{k}_1$ and $\bar{k}_2$ indicate the characteristics corresponding to the averages. These M records for groups of persons must then be used as substitutes for the N records of real persons in the testing and estimation of the relation. We may characterize this procedure as analysis after aggregation, i.e. macro analysis, as compared with test and estimation based on S_3 which may be called analysis before aggregation, i.e. micro analysis.

There are, however, more basic differences which should be noted. First, computations based on the N records of S_3 will usually give the basis for a more precise conclusion than those based on the $M < N$ records of S_5. Second, the use of S_5 to test and estimate $x_{i1} = f(x_{i2})$ requires that the relation satisfies the identity

$$\Sigma f(x_{i2}) \equiv N \cdot f(\Sigma x_{i2}/N)$$

which will only be true for a few trivial cases. Without a statistical file system based on permanent unit identifiers we have therefore to modify the hypothesis to the behavior of a group and study this instead of the persons. This may in many cases be a bad substitute for the original hypothesis.

Another and parallel aspect is when the analysis concerns the behavior of individuals, establishments, etc. explained by a dynamic model, e.g. by a relation $x_{i1}(t) = f(x_{i1}(t-1))$. The retrieval of the necessary data set will again assume the existence of the unit identifiers to be able to link data set for different periods. If not, there is no other way than looking for a third non-changing characteristic appearing in the data sets for both time periods. We also frequently form averages for establishments belonging to the same industry groups and apply these averages to study the behavior of establishments over time. In addition to the above mentioned difficulties, we have an additional problem because the establishments may have been re-classified which makes averages a rather bad basis for analyzing the behavior of establishments.

6. Information supply

The statistical file system represents a high statistical readiness. According to whether the demand for a set of data is extensive or not, i.e. if there are many users demanding information from the same set or not, we may distinguish between general purpose and special purpose supply.

The general purpose supply will normally comprise not very detailed information and will usually be implemented by statistical publications or by other mass information media. The special purpose supply will give information satisfying individual needs by means of direct communication with the user. Special purpose information will often contain details or characteristics not of general interest.

One important form of special purpose information is the supply of register service to administrative agencies collecting data. These data may later be transferred at a low cost and completely identified to the statistical file system.

The individual service implies that special precautions have to be taken, probably by appropriate legislation for the statistical information activities, in order not to injure the individual privacy.

R e f e r e n c e s:

[1] Aurbakken, E: Technical problems of Setting Up and Maintaining a Population Register by Computer, Working Papers from the Central Bureau of Statistics of Norway, No IR 67/1, Oslo, 1967.

[2] Dunn, E. S., Jr.: Review of a Proposal for a National Data Center, Bureau of the Budget, Washington, D. C., 1965.

[3] Fellegi, I and Sunter, A: An Optimal Approach to Record Linkage. Paper presented for the 36th Session of the International Statistical Institute, Sydney, 1967.

[4] Nordbotten, S.: A Statistical File System, Statistisk Tidsskrift, No 2, Stockholm, 1966, pp 99 - 109.

[5] Nordbotten, S.: On Statistical File Systems II, Statistisk Tidsskrift, No 2, Stockholm, 1967, pp 114 - 125.

[6] Nordbotten, S.: Automatic Files in Statistical Systems, Conference of European Statisticians, Working Group on EDP, Paper WG 152, Geneva 1967.

[7] Ohlsson, I.: Merging Data for Statistical Use, Paper presented for the 36th Session of the International Statistical Institute, Sydney 1967.

S o m m a i r e

Cette discussion fournit un résumé des idées, des conceptions et des principes ayant rapport à l'établissement et l'emploi d'un système d'archives statistiques. L'idée fondamentale d'un tel système est la supposition que les données recueillies peuvent disposer d'une valeur d'utilisation de longue durée, pourvue qu'elles sont systematisées et retenues de telle façon qu'elles peuvent être retrouvées et réunies facilement en ce qui concerne les diverses domaines statistiques et les periodes différentes. On discute, les exigeances qui doivent être posées concernant l'arrangement du système d'archives, la question comment le cours d'emmagasinage peut être organisé et quelles conséquences ce système pourrait motiver regardant la collection et le traitement des données et aussi les nouvelles possibilités analytiques émergées de ce système. Les aspects pratiques et les constatations faites seront traite dans une autre discussion par Monsieur Ingvar Ohlsson.