UNITED NATIONS

ECONOMIC AND SOCIAL COUNCIL

Distr.
GENERAL

E/CN.3/516
5 June 1978

ORIGINAL: ENGLISH

STATISTICAL COMMISSION
Twentieth session
20 February-2 March 1979
Item 7 of the provisional agenda. Social and demographic statistics
(b) Organization of integrated social statistics

METHODS OF COLLECTING, ORGANIZING AND RETRIEVING SOCIAL STATISTICS TO ACHIEVE INTEGRATION

Report of the Secretary-General

SUMMARY

The present document responds to a request by the Statistical Commission at its nineteenth session for a continuation of work on the integration of social and demographic statistics. The technical report, Studies in the Integration of Social and Demographic Statistics, brings together the work that had been done in this area since the publication of Towards a System of Social and Demographic Statistics. The present document discusses the next phase of this work. A draft was considered by an expert group which met in March 1978, and the present document reflects the group's comments.

Chapter I suggests action the Commission may wish to take. Chapter II discusses the changing nature of both the demand for social and demographic statistics and the technology of producing them, as they affect the possibility of and the need for integration. Chapter III explores the meaning of integration, in terms both of the formal framework and of practical problems of management. Chapter IV discusses various types of data bases and the problems of constructing them, and chapter V contains a brief conclusion. Examples of country experiences in the construction and use of social and demographic microdata bases are given in annex I and a bibliography in annex II. Annex III discusses future work, as proposed by the expert group.

CONTENTS

		Paragraphs	Page
INTRODUCTION		1 - 3	3
I.	ACTION BY THE COMMISSION	4	4
II.	THE SETTING OF THE PROBLEM	5 - 15	4
	A. Changing data needs and statistical technology: the changing nature of statistical activity	5 - 11	4
	B. The role of international discussion	12 - 15	6
III.	THE MEANING OF INTEGRATION	16 - 64	8
	A. A consistent conceptual framework for social and economic data	16 - 33	8
	B. Practical problems in the management of integrated social data	34 - 47	14
	C. The dimensions of integration: types and sources of social data	48 - 64	19
IV.	THE CONSTRUCTION OF A DATA BASE: TECHNIQUES OF CREATING, STORING AND DISSEMINATING MICRODATA	65 -102	24
	A. Single-source data bases	69 - 75	25
	B. Aggregate linking: the construction of macrodata bases	76 - 77	28
	C. Composite data bases	78 -102	28
V.	CONCLUSIONS	103 -108	36

Annexes

I. EXAMPLES OF COUNTRY APPROACHES TO THE INTEGRATION OF SOCIAL AND DEMOGRAPHIC DATA THROUGH THE CONSTRUCTION OF DATA BASES

II. BIBLIOGRAPHY

III. FUTURE WORK

INTRODUCTION

1. Work on the development of a methodology for systematizing and integrating social and demographic data has been under way in the United Nations and other international organizations since the late 1960s. During the first part of this period, a series of methodological papers were considered by various working parties and expert groups, as well as by the Statistical Commission; this phase of the work culminated in the publication of a technical report, Towards a System of Social and Demographic Statistics. 1/ This was followed by a period of less emphasis on system-building and more on the potential uses of the data and an approach towards implementation. The work of this second phase is brought together in a technical report, Studies in the Integration of Social and Demographic Statistics. 2/ This last publication contains a history of the working parties and other groups that have met to consider various aspects of the question, and a bibliography.

2. The shift in focus from purely theoretical system-building to problems of implementation has drawn attention to a number of questions not fully considered in the earlier discussions. The Working Group on the System of Social and Demographic Statistics of the Conference of European Statisticians in April 1976 discussed earlier drafts of some of the papers which are now being published in Studies in the Integration ... and concluded that what is needed next is the accumulation of practical experience in countries in working towards the integration of social statistics. To this end, the Working Group asked for a status report, which would summarize the objectives of the work and the problems that must be solved in its implementation and describe the work that has been done to date. The emphasis was to be on implementation: how to go about collecting, organizing, storing and disseminating systematized social data. The call for such a study was reiterated by the Statistical Commission at its nineteenth session in November 1976, with the suggestion that it be considered by an expert group. 3/ Such an expert group was convened in March 1978 to consider a preliminary version of the present document.

3. The present document reflects the views of the expert group. Chapter I suggests action the Commission may wish to take. Chapter II sets the problem into perspective. Chapter III discusses the meaning of integration, in terms of the conceptual framework, the kinds of practical problems - political and organizational - that must be met and the types and sources of social data. Chapter IV discusses techniques for collecting, storing and disseminating integrated social data. Chapter V contains a brief summary and conclusion. Annex I describes specific examples of various country approaches to implementation, annex II contains a selected bibliography and annex III outlines a proposed work programme.

1/ United Nations publication, Sales No. E.74.XVII.8.

2/ Studies in Methods, Series F, No. 24 (United Nations publication, to be issued).

3/ Official Records of the Economic and Social Council, Sixty-second Session, Supplement No. 2 (E/5910), para. 83 (b).

I. ACTION BY THE COMMISSION

4. The Commission may wish to:

(a) Express its views on the approach to the integration of social, demographic and economic statistics presented in the present report;

(b) Consider whether the document should be published as a technical report incorporating the Commission's comments; and

(c) Comment on the work programme proposed in annex III, especially with regard to priorities.

II. THE SETTING OF THE PROBLEM

A. Changing data needs and statistical technology: the changing nature of statistical activity

5. Traditionally, governmental demand for data (whether social or economic) has arisen from two sources. For planning, overview and general policy formulation, the demand has been for a broad brush: summary, aggregated statistics. In the economic field, it was this demand that led to the development of the national accounting framework. Although there is no comparable comprehensive framework in the social statistics area, there are partial structures in a number of subfields, and it is such aggregated data that have generally been used for policy analysis: school enrolments, aggregated up to regional and national totals; mortality rates, summarized over broad groups; aggregate numbers of families in poverty, by major demographic groups. And it is to the production of these aggregated data that the efforts of statisticians have primarily been directed. For regulation and administration, on the other hand, it is microdata - information about individual cases - that is needed: whether baby A has received a polio shot, whether pupil B has passed an exam, whether criminal C has served a sentence. The traditional uses of data and the corresponding sources from which they are derived have led to fragmentation, which is exacerbated in the social area by the lack of comprehensive theoretical framework. Producers of administrative data have had neither the incentive nor the means to relate their data to that of other producers. Compilers of aggregate data, faced with a bewildering array of often incompatible data sources, have had a tendency to ignore past work and seek new data for each new need. For reasons of cost and technical feasibility, such aggregate data commonly have given little or no information on behaviour or on distributions. The microdata, on the other hand, are often partial, biased, internally inconsistent and impossible to reconcile either with the aggregate data or with related microdata. The aggregated data are similarly likely to be incompatible from source to source and from field to field.

6. In recent years two major technological factors have been altering the supply side of this traditional picture. On the one hand, the rapid development of survey methodology has had a major impact on the way data are collected. On the other

hand, the computer has changed the way data are processed, stored and disseminated. At the same time, demands for new kinds of information have arisen as policy makers and other users are increasingly faced with questions that cannot be answered with traditional kinds of data.

7. The development of sampling techniques and the improvement in survey methodology have made it possible to obtain reliable data much more cheaply and quickly, using samples and innovative collection techniques. There has consequently been a great increase in data availability, and as sampling theory has developed and expertise in collection techniques has accumulated, the data gathered in this way have become much more reliable. The impact of this development is not confined to the statistically more developed countries; it is increasingly being recognized that a survey capability is of major importance for developing countries, where it may provide the best source of social and demographic, and often also of economic, data. It is essential that, as such a capability is developed, care should be taken in its organization and design so that the maximum potential usefulness of the data it yields will be preserved. "Progress report on the National Household Survey Capability Programme" (E/CN.3/527) is also before the Commission.

8. The rapid advance of computer technology has revolutionized the traditional approach to data processing, retrieval and general access, making possible a new conception of statistical processing and organization. Before computer technology reached its present level, the main aim of data processing was, of necessity, the reduction of the quantity of information to manageable proportions. The first step was therefore always aggregation. This meant, usually, the production of pre-specified tabulations. Once the tabulations were made, the original data held no further interest and were commonly discarded. Now, the emphasis has shifted from the production of pre-specified tabulations to the processing of primary data. Editing, cleaning and correcting can be done at the microdata level - the level of the original observations - and preservation of the microdata is both feasible and efficient. This does not mean the abandonment of pre-specified tabulations, of course: it is still as necessary as ever to prepare summary tabulations. But new possibilities for use of the data collected have been opened up. Again, this is a development that is not confined to countries with advanced technology. The new techniques are often simpler and more efficient and are helpful even at the earliest stages of statistical development when the aim is still only to produce pre-specified tabulations.

9. The changing emphasis of policy, from an exclusive concern with aggregate output to questions of distribution, and to social as well as purely economic aspects of the well-being of the population, has increased the need for this kind of information. The increasing policy importance of social programmes leads to a demand for information on interrelationships between government and households, on the size distribution of income, on the position of specific social and demographic groups and on distributions by region and type of community. Data are required for the setting of goals and the evaluation of performance in such fields as manpower training, health care and education and, to serve these needs effectively, the costs and benefits of specific programmes must be set in the context of the broader social and demographic data. These needs cannot be met by disaggregation of the aggregate tabulations alone. The emphasis is shifting from total output and

income to a concern with equity and distribution. Policy makers need to know both what the distribution of welfare and income is and how it is affected by specific governmental programmes. To an increasing extent, programmes are evaluated in terms of their impact upon different social and demographic groups. The possibility of access to microdata has, moreover, led to a change in the perceived need for this kind of information. Where good quality and conveniently accessible microdata are available, planners and policy makers do use them. The development of computer technology, furthermore, has made microanalytic modeling feasible, and this technique is increasingly used by both researchers and policy makers to simulate the distributional effects of changes in programmes and to estimate their costs.

10. The sorts of data needs that are emerging emphasize a new concern with the relationship between the microdata and the aggregated data within which they fit. To evaluate a specific programme, for instance, it is useful to be able to compare recipients of its benefits with non-recipients, and to measure the impact of the programme in terms of the recipients' whole life situation. Does a new health programme actually reach the client group it was intended for? How do the social and demographic characteristics of the recipient group compare with those of the population in general? Is the programme significant in terms of its impact upon the people it does reach? What are its costs, relative to measurable benefits? These are questions that cannot be answered with either aggregate data or microdata alone.

11. The changing technology and the changing view of the needs of society for the production of detailed and easily retrievable statistics have clear implications for the organization of statistical work. This is not the main topic of the present document and it has been discussed in detail elsewhere; reference might particularly be made to The Organization of National Statistical Services: A Review of Major Issues. 4/ Nevertheless, there are some aspects of the problem of organizing data where the bureaucratic organization of the work has a critical bearing. This is especially true with regard to co-ordination and in connexion with the questions of confidentiality and access. Where the organization of the work is an essential consideration, it will be mentioned.

B. The role of international discussion

12. The general discussion of social statistics on the international level has, until now, been conducted largely in terms of the traditional view of the statistical function and the traditional technology. It has been mainly a discussion of what tabulations should be produced. The more recent work does not propose specific tabulations, but this is a consequence more of the difficulty of the task than of a change in viewpoint. Instead of a particular selection, the national statistician has been offered a wide variety of possible tabulations, reflecting the multiplicity of attributes of social data and the impossibility of considering them all simultaneously in a cross-tabulation. But the discussion has

4/ United Nations publication, Sales No. E.77.XVII.5.

been couched in terms of aggregate data, and it is among aggregated data in different social fields that links have been sought. There is, in fact, little in the discussion to date that would have been technically unfeasible 20, 30 or even 40 years ago.

13. This is not nearly as true of the discussions within individual fields. In the demographic area, for instance, the recommendations for the 1980 census programme advocate a microdata base approach to the preservation of basic census data as well as the preparation of specified tabulations.

14. There is a need to bring the over-all discussion up to the level of the most advanced present practice. By investigating what countries are in fact doing and planning to do, it may be possible to anticipate the probable direction of development. It is with this question that this report is concerned. That does not mean, however, that the report is irrelevant for countries not at the frontier of statistical development. There is no need for countries newly embarking on work in this area to repeat the whole process of development through which the statistically more advanced countries have gone; the lack of established practices may even remove an obstacle faced by the more advanced countries. Like most technological revolutions, this one has made the task simpler, not more difficult. Co-ordination at the planning stage and the establishment of consistent definitions and classifications from the beginning of statistical work will greatly increase the productivity of statistical work at any level of development. At the same time, it bears repeating that what is being discussed here is a relatively new concept, and it should be looked upon as a guideline or goal for the future, to be implemented gradually so as to cause a minimum of disruption to ongoing programmes. The principles are simple and can be used as a guide in new work; but the realization of their full potential must be a much longer-run objective.

15. The main text of this report synthesizes the conclusions derived from the examination of country plans and their implementation. While there are some references to specific examples, most of the discussion of actual examples of country practices is contained in annex I.

III. THE MEANING OF INTEGRATION

A. A consistent conceptual framework for social and economic data

16. A consistent conceptual framework is a necessary but by no means sufficient pre-condition to the development of integrated social statistics. It was to the development of such a framework that most of the prior international discussion has been addressed, and much progress has been made. The content of social statistics has been identified and classified, fields of social statistics delineated and the kinds of information originating in each field laid out in some detail. The basic concern of social statistics with households, families and individuals has been established, and the attributes that are common to more than one field of social statistics have been identified. The fundamental importance of standardized definitions and classifications has been recognized. There is no need to repeat here what is by now well established. Some points, however, have not been sufficiently stressed and some logical conclusions not sufficiently followed through. The conception of interrelationships has been somewhat simplistic, and the impact of the new technology has not been adequately taken into account. Also, more attention needs to be paid to the relation of social and demographic data to other kinds of information, particularly economic data. Social data do not exist in an isolated world; they touch upon economic concerns at many points. Income and occupation, for instance, are both social and economic variables; and all social programmes have budgetary implications of some sort, either public or private. Furthermore, data management problems are to a large degree similar across subject-matter areas, and lessons learned in the areas where integration has proceeded furthest can profitably be transferred.

1. The relation of microdata and aggregate data

17. The first element of the conceptual framework that needs more emphasis is the necessary, logical relationship of microdata and aggregate data. Both these terms are, of course, relative. Most microdata can be further disaggregated (i.e., the household to the family to the individual to particular aspects of the individual). Most aggregate data can be further aggregated (the class to the school to the town to the region to the nation, or the first stage of the first level of education to the whole first level to education as a whole). The data spectrum is a continuum, from the smallest individual unit to the largest aggregate. In this report, the term "microdata" is used to mean data at the level of disaggregation at which they were originally collected and containing all of the substantive information that was collected. The term "aggregate" is used to refer to any summarization of microdata derived by adding up similar units. But it is obvious that what is microdata in one context may be regarded as aggregate data in another. While in the social and demographic field the reporting unit is most often thought of as the individual or household, this is not always the case. It may be the school, the school district etc.

18. Microdata and aggregate data should be viewed as complementary, not as alternatives. Conceptually, aggregate data can only be derived by combining microdata, though in practice estimates are often made by taking advantage of observed regularities in aggregate relationships, so that the necessary link

between the individual data and the summary totals is lost. This is an obvious truism where the aggregate is a simple summation, i.e., total consumer expenditures. But it also applies where the derivation of an aggregate construct involves more complex arithmetic, as for instance in the estimation of life expectancies. What is important is consistency between data at all levels of aggregation. The analysis of microdata needs to be set in the over-all framework of the aggregate data and the aggregate data, in turn, need to be supported by microdata. For any given aggregate construct, it is not difficult to conceive of an underlying set or sets of microdata. The microdata, when properly weighted, should yield that construct. It should be possible, for example, to construct a microdata set containing data on household income by type that would sum to the aggregates shown in the national accounts. This does not mean that, in constructing such a microdata set to underlie an aggregate, the burden of adjustment will always fall on the microdata. Rather, the process should be one of mutual accommodation, where both conceptual and statistical adjustments may be needed on both sides. In cases where the aggregate data have not actually been derived by aggregation of microdata, the aggregate concepts may be difficult to reproduce at the micro level, and in such cases adjustment of the aggregate concepts may be called for. At the same time, by combining data from different sources, macrodata may achieve a higher level of reliability in some areas than is possible from a single microdata source. Pursuing the household income example, for instance, it is well known that some types of income are more easily obtained from surveys, and other types from administrative sources such as tax records or social insurance. A better total can be obtained by using all available sources in combination.

2. The importance of consistent concepts, definitions, classifications and reporting units

19. As has repeatedly been pointed out, social theory lacks a single, comprehensive organizing framework. Unlike the economic field, where money can usually serve as the unit of account and the national accounts give structure to the data, the social field does not even have a common numeraire. The concern is, assuredly, with people, but people in many different guises: sometimes as individuals, but often as families, households or members of other groupings. Furthermore, there are some aspects of social information which, although impinging on people, are not conveniently enumerated on the basis of persons as the reporting units. It may be the school, the hospital, the village or, for some questions like occupation and income, even the enterprise or establishment.

20. There are some subfields of social statistics where theoretical frameworks have been developed. This is the case particularly with demographic statistics. It may be questioned, however, whether the existing theoretical structures are appropriate for use as organizing principles for collecting and storing data. Rather, they are particular analytic techniques, suited to some uses but not to others, and data arranged to meet the needs of one technique often are not appropriate to another.

21. Some social scientists would broaden the scope of social statistics to comprehend all economic statistics as well as demographic and other social statistics. Those who take this view emphasize the importance of the integrating

structure provided by the national accounts. Jurisdictional controversies aside, it is clear that the national accounting structure can have a very important role to play in organizing social as well as economic statistics, and this report will suggest some ways in which the national accounting framework can be utilized. But the national accounts, as conventionally viewed, contain only aggregated economic data, and what is needed is integration of economic and social data at the microunit level.

22. The main instrument for the integration of social data therefore has to be consistency - consistency of concepts, definitions and classifications, and consistent treatment of reporting units, throughout the entire field of social statistics wherever the same elements occur. There are of course some definitions and classifications that are essentially unique to particular subfields, as, for instance, the classification of diseases. There are others, however, that cut across several or even most fields of social statistics. Such attributes as age and geographical location are nearly universal, and even the more specialized classifications are often of interest to related fields. Classifications of level of education and occupation may be needed, for example, in studying health, social mobility or manpower planning.

23. While there may be some conflict between general and specialized needs, it is usually possible to satisfy both if the basic principles applicable to the construction of any classification system are followed. A classification system in its smallest units should be thought of as a set of building blocks, which can be combined in different ways for different uses. Where, for example, specific classifications are mandated by law for particular regulatory purposes, the classification systems should be able to accommodate them, for only if this is done will it be possible to evaluate performance under the programme in question in the context of the social and economic situation as a whole. This does not mean that all such existing mandatory classifications and concepts are equally useful; rather, they are part of the social fabric which must be taken into account in any analysis of behaviour. Thought of in this way, consistent classifications and definitions will be found to have advantages for specialized users as well as for those wishing to combine data from different fields. The building-block approach is particularly important when it is recognized that the law or regulation giving rise to the mandatory classification may change, or when the objective is to study the impact of proposed changes.

24. Nevertheless, it may at times be difficult to convince the proprietors of specialized data using specialized classifications of the advantages of following uniform standards, and this has organizational implications which are discussed below. But two points may be noted here. First, consistency does not mean identity. So long as classifications and concepts can be reconciled, variants can coexist. The conceptual framework should be conceived of as a family of interlocking interrelated classifications and concepts, not as a rigid Procrustean bed into which all data are forced. Secondly, the main objective should be forward-looking. If consistent, integrated classifications and concepts are developed (and thoroughly worked out), they are likely to be used - simply because it is easier to do so. Here, the experience in the economic field is relevant. Before standard industrial classifications and standard trade classifications were

developed, there were a great many different variants of each in use, often within the same statistical agency. Since standard classifications have been available, however, they are used as a matter of course in most new work. Construction of a classification is difficult, and when a standard form exists, the incentive to use it is very strong. In the course of arriving at a standard classification, some existing classifications may have to be by-passed if agreement cannot be obtained, but with time such by-passed classifications will decline in importance.

25. The need for consistency applies throughout the whole process of collecting, compiling and disseminating statistics. Links at the aggregate level are not sufficient. Integration at the micro level is essential, and this requires attention to concepts and definitions and care in the choice of reporting units. The behaviour of aggregates reflects structural shifts as well as the behaviour of micro units, and with only aggregate links the two cannot be separated. Fluctuations in the birth-rate, for example, reflect demographic shifts in the population as well as changes in the behaviour of individual women, as analysis of microdata would clearly reveal. It is often argued that ingenious mathematics applied to aggregated data will also reveal such relationships: but aggregation by its very nature hides more detailed interrelations and even complex mathematical methods cannot reconstitute information which no longer exists. Such methods of mathematical estimation necessarily depend upon assumptions about detailed interrelationships which in periods of social and economic change cannot be assumed to remain constant; the assumptions often relate to the problem under study.

3. The need for flexibility and the concept of the data base

26. As has been noted above, the traditional approach to data processing has involved the development of cross-tabulations which serve both to reduce the volume of data and to provide the input for further analyais. For international reporting purposes, such standardized summary tabulations are clearly required. But as a basis for national statistical work, this approach does not go far enough. Social and demographic data have very many attributes, and it is impossible to anticipate or to specify in detail what cross-tabulations will be needed. Cross-tabulation in depth is not a solution, because it rapidly leads to an explosion in the number of cells required. If, for instance, one wishes to consider 10 characteristics, each of which is classified into only 10 categories - both rather small numbers to characterize socio-demographic data - the number of cells in the matrix would be 10^{10}, or 10 billion. It was for this reason that in Towards a System of Social and Demographic Statistics it was found possible only to specify items of data and kinds of classifications, not proposed tabulations. It was, of course, not intended that any country would produce all of the tables implied by the data items and classifications listed; countries were expected to choose from among them those considered useful.

27. Even when empty cells are eliminated, any extensive development of cross-tabulations will generate a larger body of data than the set of original observations. Thus, for example, the cross-tabulations of the United States census of population fill many more computer tapes than do the original individual census reports, despite the fact that the data cross-tabulated seldom exceed

three or four attributes, and despite the large size of the United States population. Aggregation is thus an effective form of data reduction only when the information finally provided is severely restricted.

28. In addition, the pre-specification of specific forms of aggregation and cross-tabulation vitiates much of what it was hoped that integration of social data could achieve. Once the choices have been made and the data tabulated accordingly, the possibility of making different choices is precluded. If the data are aggregated in terms of one characteristic, they cannot be studied in terms of another. In the analytic applications discussed in Towards a System ... this aspect of the approach became very clear, in that these applications were limited to two-way interactions. But the bridge between such analytic data sets and the management of the whole body of social and demographic data was not well worked out. Any form of aggregation of necessity reduces the information content of the data. The information contained in the microdata on the joint distributions of different variables is lost. In microdata, on the other hand, all of the information relevant to a specific microunit is available as a separate and distinguishable set. In the limiting case, cross-tabulations and microdata approach one another if only identical cases are combined in one cell and all of the characteristics are preserved. But this is not the usual conception of cross-tabulation. Cross-tabulations at a summary level do meet the information needs of many users and will always be necessary, but they cannot serve all the analytic needs of all users.

29. One of the chief contributions of the computer revolution is that it offers a solution to these problems. The physical task of arranging and rearranging data is no longer so onerous as to limit to one pre-specified format the ways in which data can be summarized and presented. If the data are preserved in the form of the original microunit observations, they can be recast in any desired way whenever the need arises, permitting far greater flexibility in generating the wide variety of aggregations and cross-tabulations which may be required for different purposes. In some cases, samples of such microunit data may be released to other users with appropriate confidentiality protection, but the desirability of preserving the microdata records does not depend on whether or not there is any such intention of releasing individual records. This form of storage, furthermore, not only preserves flexibility; it in practice compresses the data storage space needed if the number of cross-tabulations desired is substantial. Thus it is increasingly clear that data are most efficiently stored in the form of microunit records relating to each separate reporting unit. This generalization is, of course, not limited to social data. It is equally true of economic, scientific or technical data.

30. Increasing recognition of the validity of this principle has led to the development of the concept of the data base. The term data base is used in a wide variety of senses, but in this report (unless qualified with the adjective "macro") it will be used to mean a single body of related microdata. Such a body of microdata can be of any extent - from one box of punchcards or one computer tape to a complete register covering many aspects of a country's whole population. Its essential characteristics are that it contains individual observations, usually at the level at which originally collected, and that it has been prepared

in such a way that it is available for repeated use. The kinds of data bases which different countries will find it useful to construct and preserve will differ. It goes without saying that the coverage of a single data base should be limited to what is technically manageable in a given statistical environment. Much of the scepticism regarding the feasibility of utilizing microdata sets has arisen in the context of projects which exceeded the technical capacity of the installations involved, resulting in high levels of frustruation and considerable cost overruns. It also goes without saying that a data base should be constructed with some well specified set of uses in view, though specific uses may change from day to day as the problems that are engaging the attention of policy makers and programme administrators change. As in any statistical enterprise, the designer of data bases must to some extent anticipate demand. But the approach differs from the more traditional ones in that, while data collection will still be addressed to meeting immediate perceived needs, the full potentiality of whatever data are collected can be preserved. The ultimate objective must be the creation of a multipurpose tool.

4. The integration of data bases

31. In volume terms, it is probably true that most users of statistical data are content to confine themselves to data bases originating from a single source, designated here "primary data bases". Such users are concerned with relatively narrowly defined questions and do not require information that is not contained in a single source. The construction of such primary data bases is the natural first step beyond traditional tabulated aggregate statistics, and it is likely always to be of great importance. One of its main uses will continue to be the more efficient production of the traditional aggregate tabulations. There will always be a demand for specialized data on specialized topics, and appropriate data bases to meet these demands will be needed. The concept of integrated social statistics does not imply any interference with such specialized data bases, apart from the consistency standards and whatever controls the society may wish to impose regarding such matters as confidentiality.

32. Nevertheless, integration of data from different sources at the microunit level can greatly increase their analytic potential. Linking of data over time makes it possible to trace the distribution of changes - in income, in education, in health, in whatever field is wanted. Such a technique has, for instance, been applied to income data in Sweden and Denmark and to social security data in the United States. Linking of data for a single time period from two or more sources can make possible broader kinds of analyses, taking more characteristics into account. In Canada, for instance, the merging of records of unemployment insurance with tax records has made possible studies on the migration of the unemployed and their pre- and post-migration income. In Malaysia, the matching of family planning programme data with birth registration data has contributed to evaluation of the family planning programme. In the United States, the merging of income tax and social security records has facilitated analysis of taxable income by age, sex and race. Other possibilities come readily to mind: the merging of data on consumer finances with data on consumer expenditures, which has been done on a limited basis in Canada; the addition of family characteristics to records of school achievement of pupils (tried in Mexico) etc.

33. The construction of composite data bases does not of course manufacture information that was not contained in the original primary data bases. There is a danger that if the construction is not rigorously done, the results will reflect only the assumptions made, and there is considerable opportunity for misuse of the composite data by users unaware of their origins. For this reason, the question is frequently raised whether actual merging of separate data bases is really needed or whether linking at some more aggregate level is not sufficient. Linking at an aggregate level is, again, a natural second step, and it is an approach that many countries have used. There are many uses to which it can be successfully applied. The question of when merging of microdata bases is justified will be discussed in more detail below. It may be pointed out here, however, that there are two kinds of considerations which apply. One is the same as that which applies to any statistical construct: it is better to do it once in a standardized, carefully worked out way even where the ultimate objective is the aggregate link than to reconstruct the aggregate link on an ad hoc basis every time a new use arises. But beyond this, there are more cogent reasons. The construction of composite microdata bases, by erecting a microdata framework onto which all kinds of information can be mapped, creates a higher potential for the production of useful statistics, adds new dimensions, makes possible new analytic studies. It greatly enhances the possibilities, for instance, for microanalytic simulation modelling, which cannot be carried out with aggregated data.

B. Practical problems in the management of integrated social data

34. Development of a consistent conceptual framework is only the first step towards the achievement of integrated statistics. Putting such a framework into place involves practical problems of major importance. In most countries there already exists a large body of social statistics which are likely to be fragmentary, inconsistent and under no co-ordinated control. The existing situation is often characterized by long-standing and highly decentralized administrative record systems. Any attempt to introduce order into this generally prevalent near-chaos must be carefully timed and phased. Each country is likely to face its own particular combinations of practical constraints, and there is not much to be gained by attempting to lay out a common step-by-step procedure. But the same kinds of problems do arise nearly everywhere, in differing combinations. Some of the more common constraints are discussed below.

1. Confidentiality and privacy

35. Concern over the confidentiality of information and the privacy of respondents is without doubt one of the most troubling political problems faced by statisticians today. From the point of view of the public, the concern is an entirely legitimate one; technological developments (not limited to computers) have greatly increased the ease with which information normally regarded as private can be obtained by persons without legitimate access to it, and at the same time there has been an enormous increase in the amount of information gathering that is generally considered legitimate. A relatively small part of the widespread invasion of privacy can be attributed to statistical data gathering. Most of it is

administrative or regulatory, and it is dangerous to individuals to the extent that it can have a direct impact on the person to whom it relates. Information on debts and debt-repayment gathered by a credit bureau, information on personal habits collected by a prospective employer, information on prior arrests kept by the criminal justice system, information on sources of income shared by the income tax department and the police: these are the kinds of invasion of privacy that the general public fears. But it is very difficult to maintain a distinction between such administrative and regulatory uses of information and purely statistical uses, either in fact or in the minds of the public. The administrative uses are difficult or impossible to control, whereas it is relatively easy to regulate or prohibit various statistical uses, and legislators have often reacted to the public concern by imposing controls on the statistical uses of data, prohibiting certain kinds of analysis or in some instances even prohibiting the gathering of certain kinds of data.

36. It is thus imperative to ensure that statistical uses do employ adequate safeguards of the confidentiality of information about individual respondents and that the nature of these safeguards be widely understood. Paradoxically, this is a problem that may become easier to deal with as integration progresses since, barring outright subversion, the danger of disclosure is greatest where control is fragmented and there are no uniform rules. But faith in inefficiency as a safeguard is widespread, and it is behind much of the political opposition to statistical integration.

37. The problem of confidentiality is thus of primary importance in designing any data system, but it is especially so where individual records are preserved in the form of a microdata base. A data base may contain identified or unidentified data. The technical utility of retaining the identification information in each individual record is readily apparent. But identified microdata do intensify problems of confidentiality and privacy. Close control is required so that access to identified microdata is limited to statistical uses. Several ingenious approaches to maintaining such control have been developed in various countries. One, which is in effect in Sweden, requires the licencing of each body of identified microdata and the auditing of its uses. Accuracy of the information is enhanced by allowing the subjects of the records to examine the data about themselves. A somewhat less stringent device along the same line has been applied to some criminal records in the United States; under this system, a record is kept of each time the file is consulted, and that record is open to public scrutiny. But devices like these are often not enough to assuage public uneasiness. In some cases, coding of the identification information may help to prevent unauthorized access. Such coding may take various forms, leaving larger or smaller parts of the identification openly usable, depending upon the particular circumstances.

38. Because of the confidentiality problem, however, it is often more practical to remove most or all of the identification information. This is in many cases an entirely feasible approach when the data are concerned with individuals or households, where the number of individual reporting units is large and there are likely to be many cases with quite similar characteristics. Such unidentified microdata bases may be made available if desired for general release, as was done

with the Public Use Sample (PUS) of the United States population censuses in 1960 and 1970. Where this is done, careful disclosure analysis is needed. The 1970 PUS, for instance, was released in six different versions each containing somewhat different items of information, on the ground that including all of the information in one data base would in some cases have made identification possible. For other types of reporting units, such as governmental entities (i.e., school districts) or large enterprises, suppressing identification information may be less feasible, since the individual reporting units are often quite unique and their characteristics are sufficient to identify them. But in these cases anonymity is also sometimes less necessary, since the information may already be in the public domain.

2. Respondent burden

39. A second question that has become a political and policy issue is that of respondent burden. As in the case of confidentiality, this is not primarily a statistical problem. Most of the forms that flow into the government ~~bureaucracy~~ bureaucracy from the private sector are required for administrative or regulatory purposes; by far the greatest number usually relate to taxation. But the repercussions of public annoyance with administrative forms affect statistical data gathering. The whole statistical enterprise is predicated upon co-operation from the respondents who supply the basic information, and it is essential to keep their goodwill. To a phenomenal extent, statisticians do have it: the trouble that individual respondents will go to in order to obtain the information requested is often quite surprising, even when response is entirely voluntary. In order to keep the situation from deteriorating, however, this factor must be taken into consideration in the design of any collection programme.

40. First, this requires making efficient use of the data that are collected. Because of the existing chaotic state of social data, there is, as noted above, a tendency for each new need to be met by seeking new data. Systematization and integration of the data collection and storage process would make it possible for prospective users to find and use what already exists. Where new information is in fact really needed, it can often be obtained by modification of an existing data collection programme rather than by the addition of an entirely new one.

41. Secondly, the need to consider respondent burden has implications for the over-all design of data collection programmes. From a technical point of view, there are great advantages to a programme of overlapping samples, such that the same respondents will be included in different surveys. Care must be exercised, however, to see that this technique does not so increase the burden on the particular respondents who fall in the sample that they cease to be willing to co-operate. Everyone's patience has a limit, and at some point the quality of the responses will deteriorate and the proportion of refusals will rise. Techniques for approaching these problems are discussed in chapter IV.

3. Access

42. The obverse of the problems of confidentiality and respondent burden is that of legitimate access. Ways must be found to permit statistical use of data without

invading privacy (or seeming to). Often, the obstacles to legitimate access are purely bureaucratic and could be solved in time by attention to the organizational structure. Confidentiality, for instance, is sometimes used as a tool for protecting a governmental agency's monopoly position or for denying information to a rival agency. In such cases, standardized access rules can be of great utility. In other cases, confidentiality is invoked to protect the data collecting organization, not the respondent. This happens with both governmental and private data-gatherers - for instance, when employers refuse to release salary information. Right-to-know laws can be helpful here. In some cases, information is classified as confidential even though it is already in the public domain. In the economic area this is sometimes true of the industrial register, but it occurs with social statistics as well. In many countries vital statistics records are an example. Perhaps the greatest bureaucratic obstacle to legitimate access, however, is simply long-standing tradition, and organizational considerations are of great importance in breaking down such barriers.

4. Organizational implications of the need for confidentiality and access

43. The objectives of confidentiality and access are to some extent conflicting, and it is essential to arrive at an accommodation between them that does not compromise either one too severely. The solution adopted in any individual country will depend upon its own circumstances, political preferences and priorities. It seems to be generally true that both problems are of less concern in developing than in developed countries. But there are some general considerations that would seem likely to be valid in most circumstances.

44. Uniform confidentiality rules for statistical uses of data should be established which are applicable throughout the statistical system, including those arms of it that are by-products of administrative functions. It is often the case that strict disclosure rules are established for data that are gathered for statistical purposes only, such as census data, but much less attention is given to the statistical data arising from administrative procedures. The administrative agencies themselves often have somewhat lax disclosure rules, sometimes exchanging information for enforcement or regulatory purposes on a rather casual basis. Whether or not this is deemed to be a legitimate practice in any particular administrative use, it should not be allowed to extend to statistical uses of the data. No unsanitized (identifiable) microdata of a confidential nature should ever be released from the statistical system. This principle is not quite as simple as it sounds at first. Primary disclosure is, in concept, not difficult to control through the establishment of computer routines to test any data released. Implicit disclosure, however, is a much more complex question. Every new bit of data released is an addition to all data previously released, and even where a given data set seems perfectly anonymous, it may be possible to use it in conjunction with some wholly different data set previously released to make identifications. Prevention of this sort of disclosure requires much more sophisticated testing procedures, and it can never be totally certain. It is possible, however, to foreclose any reasonable probability of identification.

45. These requirements can be met through the concept of the statistical enclave. Where administrative agencies engage in statistical activities, those activities

should be identified and isolated (for data-handling purposes) from their parent agencies. The statistical system should be considered to include both the central statistical organization or organizations and all such statistical enclaves in operating agencies, though the statistical enclaves would remain for budgetary and programme purposes part of their parent organizations. Standard disclosure rules and standard access rules, well-publicized and well-enforced, would go a long way towards meeting the public concern over invasion of privacy. In most countries, it is not the census that people are worried about; it is administrative data and their possible adversary use. What is needed is a way to extend the general trust in the true confidentiality of census data to all statistical data sources. There is no way to protect statistical or any other kind of data against a truly malignant government. But the statistical system must be predicated upon the assumption that such malignancy is unlikely. In the event that it did exist, control over the types of data collected and how they were used would be unlikely to rest with statisticians.

5. The importance of administrative co-ordination

46. Though fragmentation of statistical collection and storage is likely to lead to non-comparability of data from one sector to another, duplication of work, lack of access by users to essential information and excessive cost, a statistical office seldom effectively controls all of the sources of data. Much depends upon whether the statistical system is centralized or decentralized, the role of the statistical agency in administrative data development and the governmental structure itself. Where the statistical function is centralized, an attempt can be made to integrate the data it controls, and such a centralized statistical service may have access to a great deal of administrative data. But the statistical agency may or may not be able to play an effective role in the development and specification of the contents of administrative data bases. Even in countries where the statistical organization has such a de jure right, it is often difficult to exercise it effectively. Where administrative data originate in local or provincial governments, the problem is compounded in that these authorities may have neither the incentive nor the resources to follow national standards. Each administrative agency, furthermore, serves its own client group, which may see little advantage in co-operation. The administrative infrastructure and the level of ministerial co-operation play a major role in determining the effectiveness of any co-ordinating machinery.

47. On the other hand, it should also be noted that the benefits of effective co-ordination are great. Where agreements on standardization are reached, whether voluntarily or through mandatory controls, entrenched habits can be changed - not always, of course, but often enough to make the effort well worth while. When information is asked for in a standardized, systematic way, the originating organizations do (again, with time) change their internal reporting systems. The process is likely to be slow, but the rate of progress is also likely to be a function of the priority attached to integration. Where this is seen as a primary objective throughout the statistical services, more rapid progress can be made. The priority attached to such an effort is in turn in large part a function of the extent to which the benefits of integration can be articulated in terms of the specific advantages that would accrue to the policy makers and administrators in each of the concerned agencies.

C. The dimensions of integration: types and sources of social data

1. The basic sources of social data

48. Social data are derived from two basic kinds of sources: from specially designed inquiries such as surveys and censuses, on the one hand, and as a by-product of administrative processes, on the other. Traditionally, these have been considered to be and treated as if they were quite separate and distinct. Typically, statistical organizations have been primarily concerned with survey data (broadly defined to include complete census enumerations). Administrative data have frequently remained the province of the operating agency concerned, both with respect to the determination of what data to collect and with respect to their subsequent storage and control. A first step away from this dual approach often involves giving the statistical organization access to administrative data after they are collected, but without any role in their design.

49. Increasingly, however, the need for co-ordination is being recognized. The two kinds of data are complementary and the distinction between them is becoming less clear-cut. On the one hand, statistical monitoring is increasingly being used as a regulatory tool, in which compliance with a regulation is defined as meeting some statistical norm, such as the proportion of women or minorities enrolled in professional schools; the percentage of market share controlled; the number of days per month of acceptable air quality. On the other hand, administrative agencies are increasingly responding to statistical needs by altering or adding to administrative forms to accommodate purely statistical purposes. Most commonly, this takes the form of adding standard identification information to a form - either a common identification number, if one exists in the country, or enough information on name, address, age, sex etc. to make identification highly probable. But this is not the only sort of modification that is taking place. Purely informational questions are also being added, such as the questions on residence which have appeared on the United States personal income tax form. The information sought is used for an administrative purpose (allocation of revenue to local governments), but it is needed in statistical form; no decision relating to the individual taxpayers depends upon it. In some instances, also, administrative forms can be substituted for or used in conjunction with censuses or surveys. This has been done in Canada, for instance, where the income tax mechanism has been used as a source of data on unincorporated enterprises in the merchandising and service trades.

50. As experience in the analytic, as opposed to purely regulatory, uses of administrative data accumulates, it is increasingly being found that a combination of administrative and survey data can be very fruitful. A survey, in such a use, is undertaken specifically to provide information needed for evaluation of the administrative programme but not available from the administrative records alone. Such a survey, for instance, is the United States Survey of Income and Program Participation (SIPP), which is designed to identify the groups who are actually benefiting from various social assistance programmes and to measure the benefits relative to the recipients' income from other sources. The results are used in conjunction with administrative data on cost in evaluating programme effectiveness.

51. Even apart from such direct instances, it is often true that administrative and survey data can supplement each other. Social insurance, for instance, may cover some age, sex and ethnic groups better than the population census. This is especially true where old-age pensions and/or family allowances for children provide a financial incentive, but may also occur through employer reports on employees. Information on some (perhaps most) types of income is also generally more completely reported by the payers than the recipients, and for types of income such as interest, dividends and other transfers where household surveys yield very incomplete information, use of administrative sources such as business tax records or social security benefit payment records can be very helpful. Conversely, survey data can provide details on distribution and supplementary attributes not given in the administrative data, and where administrative sources are unreliable or inaccurate, surveys can be more closely controlled.

52. Rather than viewing survey and administrative data as alternatives, therefore, it is better to view them both as part of the cumulative addition to knowledge, which should be used in whatever way is possible to meet the needs of the society for information, subject to whatever constraints may be placed on their use for the protection of privacy or similar reasons. The society operates on its information base, both for management of day-to-day operations and for longer-range planning. It is apparent that regulatory and administrative data and survey data overlap. They often cover the same entities, sometimes from almost identical points of view. For the integration of social statistics to be considered complete, both sources of data must be included and their coverage of similar topics brought together.

2. Census and survey data: characteristics and types

53. There are a great many different sorts of censuses and surveys, and from the point of view of data integration the different sorts pose rather different kinds of problems. The most obvious is perhaps the difference between a sample and a complete enumeration. One can, at least in principle, expect a census to be complete; every individual who falls into the categories covered should be traceable in it. Where two censuses cover the same categories, it should be possible to find the records relating to a particular reporting unit in both censuses and match them. Such a procedure was followed, for instance, in Canada where information on a sample of farm operators gathered in the 1971 agricultural census was matched with information on the same farm households gathered in the population census. Since both were complete enumerations and both contained the same identification information, 98 per cent of the cases were matched successfully on the first computer run.

54. Samples, on the other hand, pose quite different problems. A single survey yielding all of the desired kinds of information relating to any given group of respondents is impractical. Any given survey is limited both by cost and by the reporting burden on individual respondents. There is necessarily some trade-off between sample size and questionnaire size. In complete censuses, the number of items that can be obtained from every respondent is limited; more detailed questions are asked only of samples of the population. Even these samples, however,

are very large when compared with those used for surveys which seek to obtain extremely detailed information about each respondent, such as a survey of consumer expenditures. The probability of finding the same reporting unit in two different samples, even when they cover the same category of respondent, is equal approximately to the product of the sampling ratios of the two samples, and where these are small it approaches zero. Various techniques have been devised for raising this probability.

55. Panel surveys represent one such approach. The same panel of respondents is continued from one inquiry to the next, so that it is possible to obtain a continuous record for each respondent. The Panel Survey on Income Dynamics conducted by the Michigan Survey Research Center is an example of such an approach; it now covers a period of seven years with essentially the same group of respondent households. The group is, of course, not exactly the same; some households disappear and some household members split off to form new households. The objective of the panel approach, however, is to maintain the links throughout the whole period covered by the survey.

56. Variants of the panel approach which impose less burden on the respondents are also possible. One such is a rotating panel, wherein in each new round of the survey some fraction of the panel is replaced. Each respondent, thus, remains in the panel for a limited time, but continuity is still maintained. A more complex version of a rotating panel is employed by the United States Current Population Survey. There, each member remains in the panel for four months, is dropped for the succeeding eight months and then is included once again for four months. Thus for each respondent there are observations exactly a year apart.

57. Use may also be made of overlapping samples where the interest is in connecting information on different topics at one point in time. Where the entire survey programme can be effectively co-ordinated, it can be conceived of as a nested set. No single respondent in such a set would be included in all subsamples, but enough overlap can be maintained among all of them to permit the establishment of connexions. Such a system of interrelated samples, furthermore, might be expected to go a long way towards reducing problems of inconsistency of definitions, lack of comparability of reporting units, incompatible classifications and misalignment. Kenya's National Integrated Sample Survey Programme is an example of this approach.

58. Alignment of sample survey results is a major problem. Before any survey can be considered to be representative of the nation as a whole (or any specified part of it), it should be checked against whatever aggregate data and other survey data are available. As a general proposition, no single survey can be expected to produce aggregate estimates as valid as those that can be constructed using information from multiple sources. There are many reasons for this. Some are reflections of the consistency problem - definitions, classifications, reporting units that are not quite the same from one source to another. Some can arise from weighting problems, if different weighting systems are used (e.g., individuals in one case, households in another). The major factor, however, is likely to be

non-sampling errors - responses that are simply wrong. Some types of such non-sampling errors are well known, such as age heaping or the tendency to underreport certain forms of income. Whatever the source of the discrepancy, it is essential to track it down and if possible provide a reconciliation with the best available control totals. Only then can the biases of any particular sample be appropriately evaluated.

3. Administrative data: characteristics and types

59. Administrative data may be an inexpensive source of data for the statistician, because they arise as a by-product of ongoing activities paid for out of other budgets. They are abundant, since they are required to be supplied for their administrative uses on regular repetitive schedules. Their coverage of the category of respondents to which they relate is often comprehensive or nearly so, since administrative regulations usually apply to all units meeting some criterion. On the other hand, they are seldom quite what the statistician would like to have, and may therefore turn out to be very expensive to use. Their administration is often highly decentralized, and concepts, classifications and coverage often differ from one source of data to another, or from one local area to another. Geographical detail may vary among political administrative units. The problem in utilizing administrative data for statistical purposes is to take advantage of its abundance, without becoming overwhelmed by irrelevant and inconsistent detail.

60. Standardization - i.e., consistency - is indispensable. As noted above, such standardization can benefit the main administrative purpose as well as the derived statistics, making it possible to relate one programme to another, to establish targets and evaluate performance in a wider setting and to bring related information to bear on the administration of a particular programme. Where the goal of standardization is given high priority, its benefits are often apparent to the producers of administrative data and their co-operation can be secured. The importance of the co-ordinating machinery cannot be overstressed. It is especially important that good communications be maintained among all interested organizations at the working level, where the value of standardization is perhaps most apparent.

61. The problem of co-ordination is particularly acute where the original sources of the data are local government authorities, as for instance with information on school enrolment or locally administered social assistance. When such governments collect the data primarily for their own use, they often consider the compilation of national data a quite secondary objective. Even with the best of intentions, furthermore, data arising from many local sources will present problems. Once again, the co-ordinating machinery becomes very important. Where control lies outside of the statistical organization and the number of local authorities involved is substantial, the difficulties may prove to be so great that an alternative approach will have to be found. In such instances, an alternative data collection effort under the control of the national statistical organization may prove to be the most satisfactory and least expensive method of procedure. But in cases where the statistical organization can control the design of the collection effort and can provide guidance (and, preferably, technical assistance and funding) to the local authorities involved, it may well be possible to achieve the required degree of standardization. The feasibility of

utilizing such locally gathered data will vary from case to case, and each must be examined individually. Often the statistician will have no choice, but where there is a choice it should be borne in mind that the objective is to minimize the cost of obtaining data of a given quality, or alternatively to maximize data quality for a given cost. The effort to co-ordinate local area statistics may, however, become so complex that switching to a national collection programme would be cheaper or more efficient. In some cases, the two methods can usefully serve as checks on each other. School enrolment data from local administrative sources, for instance, can be compared with school attendance data from national population censuses and surveys.

62. The sheer quantity of administrative data poses a problem in itself. While it is becoming technically possible to deal with whole administrative files, it is not, for statistical purposes, usually desirable or efficient to do so. It is seldom good practice, for production work, to push the state of the computer art to its limit, and for countries not on the computer frontier it is the height of folly. Of equal importance, given the state of political sensibilities in many countries, is the question of confidentiality. The notion that dossiers are being compiled on each individual in a population is an extremely sensitive one, and it does present real dangers of abuse. For both of these reasons, it is more useful generally to think in terms of using administrative files on a selective basis or extracting relevant data from them rather than routinely using whole files, though some countries consider the latter feasible.

63. One form which extraction can take involves the selection of particular identified cases. It may, for instance, be desired to match a list of identified cases in another data set. Alternatively, data for a particular client group can be extracted: children; the elderly; the poor; or more narrowly, female heads of households; workers in an industry exposed to foreign competition; inhabitants of a town with a special pollution problem; children repeating a given level of education more than once.

64. A more general approach to large administrative (or census) files is sampling. The problems of designing a sample to be extracted from a large and comprehensive administrative or census computer file are not intrinsically very different from those of designing a sample to be collected de novo. In some respects there is less freedom. Where questions arise it is not generally possible, for instance, to go back to the original respondent. In other respects, however, there is more freedom. Non-response does not occur, and techniques such as cluster sampling designed to reduce collection costs are clearly unnecessary. Where the files are identified, continuing panels can be extracted without considering respondent burden. This has been done, for instance, in Canada, where a number of departments are keeping longitudinal data files on the basis of sampling on the Social Insurance Number (SIN) in an identical fashion. It is also the basis for the United States continuous Work History (CWHS) and Longitudinal Employer-Employee Data (LEED) files, both of which are drawn from the administrative records of the Social Security Administration by selection of certain classes of social security numbers.

IV. THE CONSTRUCTION OF A DATA BASE: TECHNIQUES OF CREATING, STORING AND DISSEMINATING MICRODATA

65. It may be useful to summarize briefly at this point the various conditions discussed above that must be met in devising an integrated technique for handling social data. From a technical point of view, the chief concerns are consistency, flexibility and cost efficiency. Integration requires consistency throughout the entire range of social (and economic) statistics, in definitions and concepts, in classification schemes, in reporting units and weighting patterns, in standards of documentation. Flexibility requires preservation of options for accommodating multipurpose and unforeseen uses. Cost efficiency requires elimination of duplicative efforts and choice of the least expensive techniques. All of these technical considerations point to the desirability of a system of data storage and retrieval founded on the establishment of data bases containing microdata. But it is not only the technical considerations that must be taken into account. It is, in the first place, necessary to ensure that the specialized needs of particular sectors are met. Political sensibilities and organizational constraints impose limits that must be built into any plan for the organization of statistics. Maintenance of confidentiality and protection of individual privacy must be a major concern in the design of the statistical system. At the same time, legitimate access must be protected and the co-operation of all producers of data secured. Finally, respondent burden must be considered. These requirements lead to certain conclusions, which are elaborated upon below.

66. The data base concept is not a difficult one from a technical point of view, but there are a great many practical problems to be solved. As a first consideration, criteria are needed to determine what constitutes a single data base. On the one hand, what data are worth preserving at all? On the other, when should data from different sources be combined to form a composite data base? In principle, the answers to these questions should be found in the prospective uses of the data. In practice, however, the statistician can seldom rely entirely upon the user to make such decisions, and operating rules must be established. Cost is often raised as a primary consideration. It is true that the data base approach requires forward planning and investment in methodological development, and like many investments it may have its major impact in expanding output rather than in reducing total cost. For this reason, as well as for the organizational reasons noted above, a sequential approach is likely to be most successful. A start should be made where the scope of a data base can be easily defined, where it will be of obvious use and where the necessary methodological development cost can most readily be justified. In several countries, population census data have been considered the logical place to start. This topic is discussed in "Draft principles and recommendations for population and housing censuses" (E/CN.3/515/Add.1), also before the Commission. In other countries, the development of a data base for studying income distribution has been considered important. Where a continuing household survey programme is in operation, it provides a very good starting place. The question of constructing composite data bases comes much later; until the techniques of handling primary data bases have been mastered the likelihood of success in the more complex area of record linkage is small. Nevertheless, it is important that from the very beginning the

ultimate objective of across-the-board systematization should be kept in view and planning in different areas co-ordinated. If this is not done, future development will be made much more difficult and costly, and in some areas may be foreclosed entirely.

67. Systematization and integration are just as important on the technical level of file structure, editing procedures, programming etc., as on the conceptual level of definitions and classifications. Development of generalized editing programmes and modular programming can greatly reduce the cost of handling any particular body of data. Standard archiving practices can ensure that data remain accessible for future use. Every data base should be accompanied by a data description that will permit someone unfamiliar with the file to use it. Such a description is a one-time investment for each file, which can be used for any subsequent retrieval. In the existing situation, data are often poorly documented so that re-using them becomes extremely difficult, and what data are actually preserved is more a matter of chance than design. Proper documentation is a time-consuming task and, since it benefits primarily the next user rather than the originator of the data, it is often the victim of shortage of resources. For this reason, also, co-ordination and forward planning are essential, and allowance for proper documentation must be included in all budgets. Such technical co-ordination and documentation can be greatly facilitated if made the specific responsibility of a data administrator.

68. It is worth emphasizing once more that the benefits of systematization and integration, on both the conceptual and the technical levels, do not accrue only to the technically advanced user. At any stage of statistical development, the data that are gathered will be more useful if they are conceptually consistent. The process of production of statistics will be more cost effective if standardized procedures are followed. Realization of the full potential of the data base approach in terms of multipurpose use, flexibility and analytic sophistication may be a long-term goal, but the return in terms of usefulness of tabulations, quality of data and efficient use of resources is much more immediate.

A. Single-source data bases

69. The essential distinctions between raw data and a workable data base have been outlined above. Both may contain individual observations at the level at which they were collected, but a data base will have been prepared in such a way that the data are accessible for further use. It is edited and reconciled with appropriate control totals. It is supplied with information on edits, imputations and weighting, and with disclosure controls. It is evaluated in terms of the reliability and quality of the data. Single-source microdata bases are becoming increasingly common, and a number of examples are described in annex I. The general problems common to all of them are discussed below.

1. Editing and documentation

70. To be analytically useful, the data must be cleaned, corrected and edited, and this must be done at the level of the individual response. It has long been

recognized that different microdata sets may contain conflicting or inconsistent information, but the magnitude of this problem was not fully appreciated until the microdata began to be used themselves for analysis. While it has always been customary to check the individual reports, as long as the microdata were used only as the basis for aggregative tabulations, some corrections and adjustments were made at the aggregative level without carrying them back to the microdata records. Two of the most important kinds of edits are consistency checks and imputations for missing data. Glaring inconsistencies or impossible values - for example, a seven-year-old girl with 10 children - cannot be accepted. The problem of missing values also has to be faced. Many techniques have been developed to impute reasonable values for missing data. If adjustments for these inconsistencies and missing information are made at an aggregate level without a subsequent feedback of actual edits and imputations to individual records, then the creation of any aggregate at any time in the future will also have to involve making the adjustments again. Not only will this make data processing clumsy and time-consuming, but it will result in tabulations and analyses whose consistency with one another is far from certain.

71. This does not mean, however, that the original data should be destroyed. Users should retain the option of using different techniques, and this requires that the original observations also be preserved. Retaining the original data in each microrecord as well as giving the edited data, furthermore, is often the simplest way of documenting the editing procedures.

72. Editing and documentation should be carried out by the statistical agency responsible for the data collection, and the documentation should be regularly supplied as part of the data base. Such documentation must include computer specifications, of course; in addition, to make the data base useful to subsequent users, documentation must include what may be called statistical specifications. These include the content and format of the data, definitions of reporting units, classifications, permissible value ranges, weights, imputation procedures and editing specifications. Careful archiving is also necessary; tapes deteriorate and they get lost. The documentation gets separated from the data. When computers change, the technical specifications also change. In making the database approach workable, the data librarian plays a very important role.

73. The analysis and evaluation required to establish editing principles are an essential part of the statistical agency's job and should not be passed on to users to do for themselves. Revisions should not be carried out mechanically, but the causes of discrepancies should be studied and corrected at the individual-record level. Such a strategy is a vital aspect of the statistician's continuing efforts to learn about the sources of error in the data, and it is also a vital part of the product that users have a right to expect from statisticians. Editing should, of course, not be looked upon as a substitute for rigorous collection standards. The ideal strategy involves the enforcement of sufficiently rigorous collection standards so that the impact of any reasonable editing strategy is relatively small; when it is not, some remedial action is called for - re-interview, re-edit, revise concepts. For this reason as well as others, editing and correction can best be carried out immediately after the data are collected. Not only is this the time when the survey organization is most aware of the data

problems but, in most cases, it is the only time when a further contact with the respondent may be possible. This does not mean, of course, that errors discovered later should not be corrected or that new information should not be brought to bear. But the main editing should be carried out as soon after data collection as possible.

2. Reconciliation and evaluation

74. For general use, a data base should be reconciled with appropriate control totals and evaluated in terms of the reliability and quality of the data. Even after the elimination of inconsistencies and allocation for non-response, the data in any single microdata set are often very different from data derived from other independent sources. Where different sources give different results, a careful evaluation is needed to determine the causes. There may be no one best value, and it is necessary to decide which source is superior for particular uses. Once again, this analysis should be done by the collecting agency, since it is best able to do it and it is an essential ingredient in improving future work. An effort must be made to correct the biases in each microdata set to align it with the sources judged to be most accurate. As with any of the outputs of a statistical agency, releasing microdata that are known to be biased is irresponsible. Special resurveys, audits and small exact matches of records may be found useful in some cases in analysing the types of bias involved in particular data bases and in suggesting techniques for introducing appropriate adjustments. In some cases, adjustments can be based on internal relationships in the data themselves.

Corr. 1

75. Estimation is likely to be an essential part of the creation of a microdata base, just as it is in the creation of a macrodata base such as the national accounts. This should not be surprising, nor should it be considered a reflection on the quality of the data, but it is the responsibility of the statistical agency to make the necessary estimates and to explain the methodology used. Where survey data are known to be deficient, using the raw figures without alignment will only lead to biased results. Statistical agencies are often (understandably) reluctant to take on this responsibility, arguing that their function stops with such technical considerations as sampling error and that questions of alignment, bias correction etc. should be left to the user. It is undoubtedly true that if statistical agencies do not make the necessary adjustments, some users will - no knowledgeable user will accept conflicting information. But the adjustments users make will depend upon their own resources and expertise, and they will get results of widely varying quality. The unsophisticated user will accept the biased figures without realizing they are biased. As noted above, neither editing nor alignment requires the destruction of any information: the adjustments and imputations can be added without deleting any of the original information, and where the sources of the discrepancies cannot be precisely identified it is essential that this procedure be followed. Carrying both the unadjusted and adjusted figures should help to allay some of the uneasiness of both statistical agencies and users about the nature of the adjustments. Flexibility of use may also require, in some cases, carrying multiple values for the same cell. In some cases, a particular microdata set may yield more reliable results than are available otherwise, and the process of reconciliation may involve altering the control totals. But it is always important to identify the extent of any difference and to use all of the available information to arrive at the best possible estimate.

B. Aggregate linking: the construction of macro-data bases

76. Maintaining primary microdata bases makes it possible to preserve both the basic interrelations of data within the reporting unit and the distribution among reporting units, but there are many instances when it would also be very useful to be able to bring the diverse types of information that are available from different sources together. The traditional way of doing this has involved the linking of data at some aggregated level to form what may be called a macro-data base. This is the way, for instance, that the "control totals" referred to repeatedly throughout the present document are ordinarily derived, since few (if any) countries are yet in a position to use any other method. But linking is also possible at various intermediate levels of aggregation.

77. The national accounts are perhaps the most ubiquitous example of a macro-data base linked mainly at a highly aggregated level. A typical application of aggregate linking at a slightly lower level of aggregation is the construction of an income distribution by drawing information on different components of income from different sources. Such a procedure might rely on income tax data as the primary source but supplement it with information on pensions, social assistance and other non-taxable forms of income from the paying agents, and possibly on income below the taxable level from a household survey. This method can be used for many applications where the objective is to construct an aggregate cross-tabulation. Its primary advantage is that it can employ conventional tabulated statistics as input. Its main drawbacks are those that have already been noted: it requires rather laborious ad hoc estimation and adjustment and the product is one predetermined tabulation. Of necessity, the aggregation required reduces the information content of the data. An investment in linking at the microrecord level could produce a more flexible analytic instrument and one whose validity could more easily be tested. The effort put into editing and alignment would serve multiple uses, rather than only a single use.

C. Composite data bases

78. The next step beyond aggregate linking is merging two or more data bases to form a composite microdata base. A composite microdata base should be regarded as an analytic construct; it is directly analogous to what is done in constructing the macro-economic national accounts, except that the linking is done at the microunit level rather than a more aggregated level. As in the macro accounts, the objective is to map information from different sources onto a common framework, in such a way that conflicts among data from different sources can be examined and resolved in the light of the best available information. The methodology and the rationale for developing and using composite microdata bases are essentially similar for social data relating to households and for other kinds of data and other reporting units. Experience gained in creating composite data bases for enterprises or governmental units can profitably be carried over to household data and vice-versa.

79. It is sometimes argued that whatever information is legitimate can be extracted from the primary data bases separately, and there is a danger that the

composite data base will be used to draw conclusions which reflect only the method of merging used in constructing the composite data base. The latter danger is certainly real, and it is discussed more fully below. Nevertheless, merging, involving the matching of microrecords, is often a worthwhile statistical procedure, in much the same way as any other statistical rearrangement or manipulation. There is an increasing demand for general purpose microdata bases about households, containing more information than is available in any single-source data base, for such purposes as microanalytic simulation modeling to determine the distributive impact of social and economic policy and its cost to the government. The construction of composite microdata bases is an answer to this demand, but it is usually beyond the capability of the individual user to construct his or her own microdata base. In this sense, there are parallels between the development of composite microdata bases and that of the national accounts and input-output. These also were first done outside of statistical offices, which were reluctant to take on a responsibility for what was then regarded as a speculative tool.

80. A composite data base may be either special purpose - to meet a particular need - or general purpose - to create a multiple-use tool. In most countries where matching of microrecords has been undertaken, it has initially been to meet some special need where the characteristics required in the composite data base can be clearly specified, and this is unquestionably a sensible approach. Merging should not be undertaken as an exercise in statistical pyrotechnics; there should be some clear use in view, and it is very helpful for the prospective user to be involved in the statistical design from the beginning. (The difficulties that may arise when this is not done are well illustrated by several of the early examples cited in annex I).

81. Once the effectiveness of the composite data base as an analytic tool has been demonstrated in a special-purpose application, however, horizons can be expected to widen to encompass general-purpose composite data bases. Like the national accounts, which are a composite macro-data base, some composite microdata bases are of such obvious general usefulness as to require no specific defence. An example of such a composite general-purpose microdata base in the social statistics area would be a sample of households aligned to census demographic totals, containing a wide variety of social and economic attributes for each household and its members, each attribute reconciled to relevant national control totals. Using the population census (or a sample from it) as a frame, it would be useful to add such additional information as consumer budgets (usually collected as a separate exercise for a much smaller sample), information on education, occupation etc. Such a composite data base can be thought of as a model (in the sense of scale replication, like a model of a ship) of the population, onto which any and all available information about the population can be mapped. In other words, it is a receptacle for storing known information in an orderly and retrievable way. Several countries, including Sweden, Norway, Denmark and the United States, have developed such general-purpose household data bases, of varying degrees of broadness; some of these are described in annex I. Suitably sanitized, such a household data base would be an extremely powerful analytic

tool, both for governmental policy makers and research workers. The argument for constructing it, like the argument for constructing the national accounts, is that the result will be better and the costs less than comparable ad hoc data collection. For example, there have been many uses of the United States Consumer Expenditure Survey and the Survey of Consumer Finances which involved combining either or both with census or Current Population Survey data. Each time, the job had to be done entirely anew, and no single user was willing to devote the resources needed to do it carefully in a reusable form. In contrast, the project described in annex I to combine similar Canadian studies has produced a definitive reusable data base.

82. The construction of household composite data bases can be approached in many ways. What is generally referred to as "exact" matching may in fact be more or less exact. A match based on name, address, age, sex, marital status and similar factors, for instance, still involves some element of probability. But such a match, if done carefully, is likely to be very good. Non-exact methods of matching, furthermore, need not imply that the match is poor. The criteria employed in matching may be more or less stringent. For any given method, there are ways to measure the goodness of the match. The technical characteristics of certain steps in the spectrum of possibilities will be discussed below.

83. It is possible to reach any desired degree of assurance that two cases are correctly matched, but reducing the error rate entails a cost. It is therefore necessary to choose the range of error that will be accepted. The choice will normally depend on the use to be made of the data, because the error rate accepted will affect the conclusions it is safe to draw. Thus, there is an element of statistical matching in all so-called exact matching. There is a difference, however, in the criteria used in judging a good match. In an exact match, the emphasis is on the likelihood that the two linked records actually refer to the same unit or event. In statistical matching, it is on the reliability of the derived aggregate or other final output.

1. Exact matching

84. Exact matching through the use of a common identifier has obvious technical attractions, and there are a number of applications where this can be a very useful technique. In countries with population registers including a unique identification number and with relatively small populations, it is at least in concept possible to think of constructing one comprehensive data base including both census data and administrative data, all organized around the unique identification number. Such an approach, of course, is applicable only to complete enumerations; it cannot comprehend sample survey data. The Norwegian register system described in annex I is an example of such a comprehensive data base. This approach does, however, raise the privacy problem in acute form, and it is probably true that there are not many countries where it would be politically acceptable. As a minimum, it requires establishing beyond any shadow of a doubt the inviolability of the statistical confidentiality rules. Where populations are large, furthermore, exact matches involving entire populations do approach the practical limits of present computer technology.

85. Where matching cannot be based upon a unique identification number, information on one or more characteristics such as name, address, marital status or ethnic origin can often be used successfully as the basis for matching. However, because names as recorded in different files may vary, addresses change, even names change (as in marriage) and errors may occur in the recorded information, the use of such characteristics as matching variables will often add to the cost and reduce the accuracy of an exact matching operation as compared to one based on a universal and well reported unique identification number. Nevertheless, if proper preparations have been made in advance in designing and editing the files to be matched, the cost and errors can be held to acceptable levels. Of course, the definition of what is acceptable will depend upon the particular purpose for which the match is being carried out. If adequate preparations are not made, costs and errors can easily rise to unacceptably high levels. It was from several such earlier attempts that a degree of scepticism about exact matching arose. Nevertheless, as experience has been gained, the success of such ventures has increased. Some of these examples are described in annex I.

86. Where the files to be matched are both samples, exact matching is not likely to be a possibility, since different samples are unlikely to contain the same reporting units. As noted above, co-ordinated sample design can increase the possibilities of exact matching of one sample with another, but the potentialities of this method are limited by considerations of respondent burden. It is not necessarily less burdensome to complete two shorter questionnaires than one longer one, and the limits of the size of individual questionnaires are well known.

87. There are, however, several applications where exact matching is both feasible and extremely useful. One such is where information can be extracted from a census or other relatively complete enumeration to be attached to a less comprehensive sample. For special-purpose uses, it is often possible to find a basis for such exact matching even in countries which do not have population registers. In Canada, use has been made of the Social Insurance Number, and in the United States similar use has been made of the social security number. In a number of countries at various levels of statistical development, characteristics such as name, address and sex have been used jointly in matching operations. Several such projects are described in annex I.

88. A second and newer application of exact matching involves the collection of a representative sample of the population ~~which~~ that is specifically designed to be matched with existing administrative files, in order to evaluate particular programmes in terms of the economic, social and demographic characteristics of the population served in comparison with the population as a whole. The United States Survey of Income and Program Participation is an example of this approach. Only by analysing the participation of individuals in the various programmes can effective programme evaluation be carried out. For such purposes, an exact match between the administrative records and a representative sample of the population as a whole is very useful.

2. Statistical matching

89. Where exact matching is not a practical solution, either for technical or political reasons, other methods of combining the information in different primary data bases are needed. Bearing in mind the mapping principle, what is being sought is a method of bringing information from a variety of sources together and mapping it onto a microdata set, in such a way as to be consistent with the control totals of the macro data and all the distributions known from microdata sources.

90. The close relation between exact and statistical matching has been noted above. As the number of characteristics used to specify an "exact" match increases, and as the accepted error rate rises, it will shade imperceptibly into statistical matching; that is, the objective will cease to be the matching of identical units and become the matching of similar units. For most statistical purposes, this is an entirely acceptable procedure.

91. Statistical matching is also closely related in both theory and practice to editing and imputation for missing values. When missing values are imputed or the editing procedure suggests a new value as preferable to the observed value, what is in effect being done is a statistical match. Another case (or sometimes an average of cases) is found which is similar to the case under consideration, and the value of the missing or unacceptable item transferred from this statistically matched case to the case under consideration. The degree of similarity required varies; indeed, in some widely used techniques it is very slight. The so-called "hot deck" method simply uses the last case processed regardless of its resemblance to the case under consideration, on the grounds that even as crude a procedure as this is an improvement over doing nothing. But it is of course possible to use much more stringent matching criteria.

92. The technical problems of statistically matching surveys with surveys, surveys with censuses or either with administrative data do not really differ a great deal, except for considerations of sampling reliability and the difficulty of achieving conceptual consistency. 5/ More can be done with large and dense samples than with small and sparse ones, and the particular technique of matching that is appropriate in a given case will depend upon the characteristics of the data sets being matched. The job is easier if forethought in design has yielded consistent concepts and definitions. For matching to be possible, there must be suitable variables on which to base the match in the bodies of data to be merged. There is no way to match two bodies of data that contain no variables in common, and the match will be trivial unless the variables used for matching are closely related to the remaining variables in both data sets.

93. For any statistical matching process to be valid, a first requirement is that like things be matched. The reporting units must be defined in the same way. They should not, for instance, be census households in one data set and income tax units in the other. Similarly, the variables upon which

5/ An extensive literature on this topic is accumulating; annex II cites some of the references.

the match is to be based must be defined and classified in the same way. Income cannot be taxable income in one case and total gross receipts in the other. The matching variables must be aligned. If the aggregate amount of property income reported for tax purposes is double that reported in a household survey, two equal money amounts clearly cannot be matched. These problems of definition and alignment of the matching variables are extremely important. They may consume a large part of the energy of a matching effort, and certainly the quality of the ultimate match will depend on how thoroughly the definitional adjustment and alignment have been carried out.

94. Given two data sets both containing the same set of suitable variables for matching and both properly aligned, the matching process resolves itself into one of selecting the most appropriate units in the two data sets to be combined. Annex I describes a number of specific examples of techniques that have been employed. It is only through the consideration of such practical examples that it is possible to evaluate the usefulness and validity of the approach. The discussion here is of necessity quite general.

95. Traditional multivariate regression analysis is in effect one method of statistical matching. Information is imputed from one data set to another by setting up a multiple regression model to predict for each case in one data set an estimated value of a variable contained in another data set. For this method to be successful, it is of course necessary that the two data sets contain common variables which can serve as the independent variables in the regression equation. The validity of such an imputation is then dependent on how well the variable which is being imputed is explained by the variables which are in common. For many analytical purposes it is not necessary that the estimate be accurate at the individual observation level; it is merely necessary that it perform satisfactorily on average over the observed range of variation. If the regression fit is quite close, the substitution of the regression value for an actual value may not invalidate the subsequent analysis. This consideration is important: the objective is to create composite data sets which yield valid results in statistical applications, not to reproduce individual cases exactly.

96. The technique of imputation by regression is considerably less satisfactory, however, for transferring complex sets of information from one data set to another. Thus, for example, consider the problem of imputing consumer budget data to a national population sample containing other social and demographic information. A difficulty arises in that consumer outlays are all highly interrelated. A separate regression estimate for each outlay would produce an inconsistent budget pattern for any specific reporting unit. One of the major objectives of collecting consumer budget information, furthermore, is the study of interrelationships among the budget items - interrelationships which would be lost if each type of outlay were imputed independently. Although it might be possible to design a model which would take into account for each item of outlay all of the items which had already been imputed, thus attempting to preserve the relationships of the original sample, such a model would be extremely complex.

97. A simpler way of proceeding is to transfer complete sets of expenditure data from observations in one sample to observations in the other sample by a matching

process - in other words, to combine an observation from one data set containing budget data with an observation in another data set containing other social and demographic information. Such a technique not only retains the integrity of the sets of information in the two samples but also retains the observed variance. Most matching techniques, furthermore, also avoid the necessity for specifying a precise functional form of the relationships involved. Where the functional form is known and the data are so scattered that it is difficult to find cases that match in the two samples, regression analysis is likely to be a more valid approach, but with large bodies of data where similar cases do exist, imputation by matching will retain the distributional characteristics of the original samples and reflect the basic relationships more accurately.

98. It should be recognized that statistical matching is appropriate only in fairly dense data sets. Where there are only a few cases over broad intervals of the variables being matched, the likelihood of inappropriate matching is obvious. For this reason, the technique is not generally applicable to small samples or to those records in large samples which have unusual or extreme characteristics. It is also apparent that although the matching technique takes into account the relation between the matching variables and the remaining variables in each data set, it can say nothing about the conditional joint distributions of the variables in the two data sets that are not matched. Unless external evidence on the latter is available, the assumption is necessarily made that such conditional joint distributions are stochastic. Nevertheless to the extent that the non-matched variables are in fact correlated with the matched variables, the aggregate joint distributions will be correctly reflected. The objective in constructing a composite data base is not to develop links between unmatched variables for individual cases: that is clearly impossible. To attribute meaning to the joint distribution of unmatched variables within groups or cells is wholly illegitimate, since it is ordinarily either random or dictated by the matching technique. The rationale for making the imputations, adjustments and alignments at the microrecord level rather than at some intermediate or aggregate level is based on the need for consistency at the microunit level, so that all subsequent tabulations of the data will contain the same adjustments. As already suggested, where existing values of a variable known to be biased in one sample are replaced by new unbiased values obtained by statistical matching with another sample, both values should be preserved. All of the reasons enumerated above for storage of data in microunit form still apply; the microdata carry more information, and they are a more efficient form of storage.

99. Despite these reservations, there are instances where statistical matching has a number of advantages over exact matching. It is suitable in circumstances where exact matching is impossible, as in matching samples not containing the same reporting units, it is likely to be less costly (though this may not be true in the longer run as experience with exact matching accumulates). More importantly, it has significant disclosure implications. Since statistical matching creates microdata units which do not refer to any real reporting unit, it is likely to be much less objectionable politically than a process which brings together in the form of a dossier large amounts of information about actually existing households - regardless of the disclosure safeguards that may be set up to protect such information.

3. The role of the composite data base

100. The chief advantage of statistical matching, however, does not lie in these considerations. In these respects, it may be looked upon as a substitute for exact matching, better in some characteristics and worse in others. Where statistical matching opens up wholly new possibilities is in its capacity for creating general-purpose composite microdata bases, thus making possible a wide range of social and economic analyses capable of contributing to the better formulation and evaluation of governmental programmes and to the better understanding of the functioning of the private sector. The central characteristic of such data bases, the source of their analytical power, is that they simultaneously preserve the information needed for analysing distributions among social and demographic groups and at the same time can relate households and individuals to the activities of both governmental units and organizations in the private sector. Statistical matching has developed largely as a response to the demands of policy makers and programme administrators for information of this sort not obtainable in any other way.

101. Thus, in the longer term, it may be that such general-purpose microdata bases can provide the sought-for information framework for social and demographic statistics. Using the national population as the primary sampling frame, a microdata base can be established which groups the individuals in the population into family or household units. Statistical matching techniques can then be used to map many additional kinds of information onto these units. (Exact matching can of course also be used where feasible.) The adaptability of this vehicle as a receptacle for economic information - income, expenditures, labour force activity - is readily apparent. But it is equally suitable for receiving information on education, health, crime, migration, marriage, divorce and a host of other social and demographic characteristics. For some time to come, it is likely that considerations of technical manageability will require the maintenance of different versions of such a data base for different uses. A data base used to analyse crime victimization, for instance, might not need the detailed information on health. But so long as the different versions are consistent (in sampling frame, reporting unit, definitions and classifications), this is not a material consideration. Those matches that are needed can be carried out, with no more cost than computer time.

102. At the same time, similar data bases can be (and are being) constructed for other types of reporting units: governmental organizations (schools, school districts, highway commissions, hospitals, local authorities, national government ministries), enterprises and other private bodies. In one sense, such comprehensive microdata bases are a logical extension of the national economic accounts, in that one or more such microdata bases can be conceived of for each of the institutional sectors of the national accounts. But this does not in any way imply that their main function is economic analysis. Although the importance of their role in this use is increasingly being recognized as microanalytic simulation is more widely used as an analytical tool, even in this use it is the ability to take social and demographic characteristics into account that makes them so valuable. Neither social nor economic questions can any longer effectively be studied in isolation. Both social and economic data are ultimately about people and, even where they deal with other kinds of reporting units, it is essential to keep the ties to individuals and households clear. Comprehensive microdata bases are a powerful tool for doing so.

V. CONCLUSIONS

103. It may be useful at this point to recall the purpose of this report. It was to be a status report on systematizing and integrating social and demographic data with examples of work that has been taking place in individual countries. The emphasis was to be on implementation: how to go about collecting, organizing, storing and disseminating systematized social data. The report has examined these questions from several points of view, exploring the nature of social data and the political, organizational and technical problems that must be dealt with. Annex I summarizes a number of examples of country experiences.

104. The report is intended to be useful for countries at all levels of statistical development, and it is hoped that all can find something relevant to their own situations. There is a progression of steps through which countries are likely to go, moving from the re-usable single-source data base to co-ordination of concepts and integration of substance and finally to true composite data bases. For many users and many uses, aggregated data cross-tabulated on a rather detailed level could be very adequate. In cases where considerations of privacy might forbid microdata bases, macro-data bases of this kind might be very useful. On the other hand, it is clear that the progression to composite data bases has been hastened by a growth in demand. Users who a few years ago would have been satisfied with a set of cross-tabulations now want access to microdata.

105. The microdata base approach is not a luxury that is limited to statistically advanced countries. It can greatly simplify any country's efforts to systematize social and economic data. It is flexible and cost efficient; it does not require thinking of all possible future demands upon the data in advance, and it compresses storage requirements. But it does require a change in attitude and emphasis. The new technology, both in data collection and in data processing, has made possible a wholly new conception of the statistical function and, to take maximum advantage of the potentialities, a rather radical change in modes of thinking is required. Where statisticians were - and in many cases still are - accustomed to planning their work in terms of the production of a set of pre-specified tabulations, it is now necessary for them to reorient their thinking to deal with a new set of problems and a new set of goals. But the rewards of such a change in attitude and approach can be very great. The new tools have enormous potential for application to increasingly pressing social and economic problems.

106. Within this over-all conclusion, however, there is room for a great deal of variation. The data base approach will take time to implement and use, especially in countries with established statistical systems. It will, in particular, take time to accomplish the integration of data originating as a by-product of administrative and regulatory activity - and this is an important source of social data. A gradual, one-step-at-a-time procedure is only sensible. A start can be made in the construction of relatively simple and constrained data bases with obvious immediate utility. Annex I describes several such examples. As experience is gained, the approach can be generalized, and again annex I gives examples, though no country has yet traveled very far down this road.

107. The key to integration is consistency and standardization of definitions and concepts, classifications, reporting units. Applied to data at as low a level of aggregation as is consonant with any given country's circumstances, it will help to promote the development of social statistics as a unified body rather than many isolated fragments, even when the primary objective remains tabulated output. But concentration upon the production of tabulated output - even highy disaggregated tabulated output - does not approach the problem in the most useful way, and is likely to lead to increased cost, duplication of effort, lack of co-ordination, rigidity of output and inefficient use of both capital equipment (computers) and trained manpower.

108. The construction of composite data bases (whether on the micro or macro level) should be regarded as analysis, not data collection. As is true for the macro national accounts, the construction of composite microdata bases can be expected to involve estimation and imputation, and it should be done by technicians who thoroughly understand what they are doing. Though the initial steps in this direction are likely to be undertaken by analysts primarily interested in their own use of the resulting data base (as was also true in the case of the national accounts), this is not something that should on a long-term basis be left to do-it-yourself activity on the part of users, any more than one would not think of doing this with the national accounts. It is a part of the product that users have a right to expect from the producers of the statistics. In the long term, the construction of comprehensive composite microdata bases offers a very great potential as a device for integrating social and economic data of all sorts, and as an analytic tool.

Annex I

EXAMPLES OF COUNTRY APPROACHES TO THE INTEGRATION OF SOCIAL AND DEMOGRAPHIC DATA THROUGH THE CONSTRUCTION OF DATA BASES

1. This annex will describe, very briefly, selected specific examples of different approaches to the integration of social and demographic data that have been employed in various countries. The discussion is organized by country, since more often than not where there are several undertakings in one country they are interdependent. The discussion is not meant to be exhaustive, either in country coverage or in the discussion of activities within countries. Rather, it is intended to reflect as wide a variety as possible. It is limited, however, not only by considerations of space but also by the examples that have come to the attention of the United Nations Statistical Office.

2. It is not surprising that examples are most abundant in the developed countries and, in particular, those enjoying the most advanced computer technology. But they are no longer confined exclusively to those countries and, as such efforts as the National Household Survey Capability Programme (a progress report on the Programme (E/CN.3/527) is before the Commission) come to fruition, opportunities will open up for developing countries where they do not now exist. The discussion that follows is heavily weighted with North American and Scandinavian examples, because these are the countries where both conceptual development and practical execution have gone furthest. In the search for models, however, statisticians in developing countries should not dismiss the examples cited as irrelevant for them; it may well be true that the newer approaches can be simpler and more cost effective than the old if approached with caution and common sense and kept within the technical capabilities of the responsible organization.

A. The United States

3. It is useful to begin the discussion of country work with a consideration of the historical development of this field in the United States, since it was there that much of the early work took place, and it is perhaps also there that some aspects of the work have developed furthest.

1. Single-source data bases

4. The microdata base concept in the United States, in limited form, long antedates the development of automatic data processing equipment. Such a file was maintained, for instance, through the 1930s and 1940s for the information on individual establishments collected in the Census of Manufactures. In physical form, the file consisted of handwritten cards; each card contained a longitudinal record pertaining to one establishment, linked together over approximately a 20-year period. The file was not complete, in that it covered only the larger establishments, but it accounted for some 70 to 80 per cent of

the total value of production. The data recorded in the file were of course confidential and the file was therefore available for use only within the originating agency. As a consequence, it was used mainly for editing and checking purposes, but a few analytic studies were based upon it, dealing with such topics as industrial concentration and the cyclical variation in prices, output and employment.

5. One of the first computer-based microdata bases developed for general use outside the originating agency was the Income Tax Model, developed by the Internal Revenue Service (IRS) from a sample of individual income tax returns. The impetus for the construction of this data base was a research project undertaken by the Brookings Institution in connexion with a prospective revision of the tax law. The first model related to the year 1962 and contained data extracted from about 100,000 tax returns. It included such items as types and amounts of income, types and amounts of deductions, type of return, number of dependents and occupation. It did not contain information that could be used to identify individual cases, such as name or address. The data from each return were presented as an integral set. Thus it was possible to estimate the impact of a proposed change in the tax law by recomputing the tax for each individual case and then tabulating the results in various ways. The technique proved to be so successful that in very short order it was adopted as the standard method of revenue estimation by the Treasury Department, replacing the older methods based upon aggregated data. Production of the data base has now been routinized; it appears annually, with expanded amounts of detail in alternate years.

6. In the mid 1960s, the Bureau of the Census developed the first Public Use Sample (PUS) of the population census. This was a 1-in-1000 sample of the 1960 census (approximately 17,000 cases) and contained information on a variety of demographic, social and economic variables relating to the respondents. Among these were age, family structure, education, occupation, income, family income, ownership of various consumer durables including automobiles, characteristics of housing units occupied and limited information on migration. Once again, identification was removed, in this case including most of the geographical information. The data base was sold to the public, and in time came to be widely used as users' capabilities in handling large data sets improved with experience. But it was not only private researchers who found it useful; within the Government it was also widely used in analytic and policy studies. The convenience of its standardized, clean, well documented form offset the lack of identification information. The 1960 PUS was an after-the-fact reconstruction that had not been included in the original planning for the 1960 census. Consequently, it required a substantial amount of reprocessing of the basic data (which, fortunately, were still available in machine-readable form). It was in connexion with the development of this sample that many of the techniques for editing, cleaning and documenting of microdata were developed - and that the magnitude of these problems, which had been hidden in the aggregate tabulations, became apparent.

7. For the 1970 census, the inclusion of a PUS was an integral part of the programme from the start. On the basis of the experience with the 1960 PUS, however, it was considered that an expansion of the programme was warranted. The

sample size was increased to 1-in-100 rather than 1-in-1000, and a substantial amount of geographical information was added. In order to provide as much detail as possible without disclosure, six different versions of the sample were developed. The demographic and income information in the six versions is nearly identical, but they contain two different patterns of information on housing and ownership of consumer durables and three different kinds of geographical information. Where smaller geographical units are identified, fewer data items that might conceivably lead to identification are included, and conversely, where the geographical information is limited, the data items included are expanded. Construction of the six versions of the 1970 PUS proceeded as an integral part of the processing of the census itself. That they became available with very little delay after the aggregated census cross-tabulations testifies to the change in the conception of the data processing task that had taken place. Editing at the microunit level was now taking place for census processing purposes; it did not have to be done additionally to produce the PUS.

8. Reception of the enlarged 1970 PUS was enthusiastic and led to a demand for a similar enlargement of the 1960 sample. With the experience gained in the two previous efforts, this was considered to be a feasible project, and the 1960 data were once again reprocessed to construct an enlarged 1-in-100 sample consistent in design and format with the 1970 PUS. Planning for the 1980 population census has as a matter of course included provision for the construction of public use samples. A project has recently been approved to reconstruct public use samples, at the 1-in-100 level, from the 1950 and 1940 population censuses, even though these census records exist only as microfilms of the original schedules. So confident have both the producers and the users of such data bases become that this last project is considered to be a feasible one for a university research team to direct; using established procedures, it is not expected to constitute a significant burden for the Bureau of the Census even though its processing is likely to occur concurrently with the 1980 population census.

9. These pioneering efforts have been followed by a number of others. Among the most widely used have been the Social Security Administration's Continuous Work History (CWHS) and Longitudinal Employer-Employee Data (LEED) files, both 1 per cent samples of the administrative files containing data reported by employers in connexion with social security contributions. Some data items are continuously available in microdata form back to the establishment of the social security system in 1937. The LEED file contains linked quarterly information for individual respondents showing covered income from each employer for the period from 1957 to date as well as some demographic information such as age, race and sex, and information on the employer's industrial category and location. The LEED file has been extensively used in the study of migration and for studies of wage behaviour. Other data bases now available include the Current Population Survey (monthly, but not linked over time), and the Consumer Expenditure Survey (1962; 1972 now becoming available). Mention might also be made of the Survey of Economic Opportunity, covering 30,000 households for 1966 with a partial repeat in 1967. This was especially designed for the study of poverty and so is heavily concentrated on low-income groups. The construction of this data base was a long and arduous task, and it serves as a good illustration of the importance of involving the user

in the design process from the beginning. The data were originally collected by the Census Bureau for the Office of Economic Opportunity. They were then transferred, after removal of identification information, to a private contractor for editing and distribution. The task proved, however, to be beyond the contractor's capabilities, and the data passed through several other hands before the Brookings Institution finally acquired them. On the basis of its earlier experience with the Income Tax Model, Brookings was able to complete the editing, but there were many problems of inconsistencies, inappropriate classifications etc. that better preparatory work and more careful handling could have avoided.

2. Panel surveys

10. Panel surveys, in which the same respondents are included in successive time periods, are sometimes a convenient way to maintain linkages and achieve continuity over time. The LEED file of the Social Security Administration is in effect a panel survey, since the individuals included in the 1 per cent sample are all those whose social security numbers fit certain selection rules, which are not changed over time. Thus, any individual who ever falls into the sample will continue in it whenever employed in a covered industry. The file is arranged by individual; that is, all of the records for the whole period for one individual are brought together.

11. Most panel data that are collected as surveys, rather than as by-products of administrative activity, are of necessity much more limited in size and, usually, in longitudinal coverage. Of the rather numerous existing examples, mention might be made of the Michigan Panel Survey of Income Dynamics, which has followed a sample of 5,000 households over a seven-year period, and includes a substantial amount of demographic and social information about the households surveyed, as well as detailed information on incomes and occupations of the household members, and of the Parnes sample which contains information on wages and employment over a four-year period for a similarly sized sample. Both of these surveys were conducted by university research teams and are available in microdata form for purchase by the public. There is, however, nothing inherent in the methodology which would inhibit a similar effort wherever a continuing survey capability existed.

3. Exact matching

12. Progressing beyond the single-source data base, a number of examples of exact matches exist, a/ of varying degrees of complexity. One of the simplest

---

a/ A list of such examples has been compiled in the United States by the Subcommittee on Matching Techniques of the Federal Committee on Statistical Methodology. Many of them are reported on in the Social Security Administration series, Studies from Interagency Data Linkages. The discussion that follows depends heavily upon the papers of Radner and of Radner and Muller cited in annex II to the present document.

is to be found in studies aimed at evaluating or extending the coverage of lists such as sample frames. The Bureau of the Census has conducted numerous coverage evaluation studies in connexion with various censuses. For example, samples from 1950 census records, registered births and other sources were matched with 1960 census records in order to evaluate coverage of the latter. In such a match, the emphasis is upon the presence of the units in the files, rather than upon the relationships of the data in the two files. In an example of list construction, the Statistical Reporting Service of the Department of Agriculture used exact matching in the construction of a master list sampling frame of farms in each state. This master list was constructed from several different lists, and exact matching was used to detect duplication between and within the different lists.

13. A relatively simple but often very useful type of exact match involves the addition of data items drawn from a relatively complete file to cases in a smaller sample. An example of such a use was the addition of age, sex and race drawn from the records of the Social Security Administration (SSA) to the IRS Tax Model described in paragraph 5 above for the year 1970. For reasons of confidentiality, the match was performed in several steps. The IRS first supplied SSA with a list of the social security numbers of the cases in its sample, and SSA attached age, sex and race information to each social security number. The IRS next inserted the age, sex and race information into the income tax records and then deleted the social security number. The resulting matched but anonymous information was then made available for public release.

14. An example of a much more ambitious exact match was that carried out in the construction of the 1973 Exact Match File, a joint project of SSA and the Bureau of the Census, with the assistance of IRS. This file consists of an exact match of four sets of data. The base was the March 1973 Current Population Survey (CPS), a survey of roughly 50,000 households conducted by the Bureau of the Census. Among other items, the CPS provided data on family composition and cash income of persons, although the income data suffered from some deficiencies. In exact matches, SSA earnings and demographic data and a limited amount of federal individual income tax information, and later SSA benefits data, were matched to the CPS sample. This project represents the culmination of a long developmental period, during which a series of matches for earlier periods beginning in 1963 were undertaken, and the project was of very great importance in the development of the methodology of exact matching. In particular, it emphasized the importance of the preliminary work on data preparation, editing and alignment, the selection of matching variables and the establishment of tolerances, weights and thresholds. One of the by-products of this project is a very useful bibliography on exact-matching methodology, prepared by Hans J. Muller of the Census Bureau, which is included in annex II.

15. Representative of a different kind of exact-matching project is the projected Survey of Income and Project Participation, being sponsored jointly by the Department of Health, Education and Welfare and the Bureau of the Census. The objective of this survey is to obtain a representative sample of the population as a whole which can be used as a control in evaluating specific social assistance programmes. In addition to questions on various social, demographic and economic characteristics, the survey includes detailed questions on participation in

various governmental and non-governmental cash and non-cash income maintenance, insurance and training programmes, estimates of ownership of various major types of assets and summary estimates of net worth. The file obtained in the survey can then be supplemented by data on the cases included in it drawn from the administrative files of the aid-giving agencies, on an exact-matching basis. In this way it will be possible to examine such questions as whether the programmes are in fact reaching the target population they are designed to assist, whether they are of significant benefit to the recipients and how much overlap there is among programmes, as well as more general questions such as over-all income adequacy.

4. Statistical matching

16. Relatively early in the history of the development of microdata bases in the United States it became apparent that it would not be possible to obtain microdata bases containing all of the kinds of information that were needed for some analytical purposes on the basis either of single sources or composites constructed through exact matching. One of the earliest instances of the use of statistical matching grew out of the effort of the Brookings Institution in tax modelling described above. It soon became clear that the Income Tax Model data base was inadequate in a number of respects. It did not cover any forms of income not then taxable. Nor did it cover the rather large segment of the population whose incomes fell below the level at which filing an income tax return was required, and there was at the time a strong interest in the impact of proposed changes in the tax system on this low-income group. The Survey of Economic Opportunity (SEO) mentioned in paragraph 9 above reflected this interest. The Brookings project combined the SEO and the Tax Model into a composite data base called the MERGE file, in order to obtain better coverage of the whole range of income distribution than could be obtained from either alone. The actual merge involved only the middle-income groups. Those households in the SEO sample whose incomes were below the tax-filing limit obviously could not be matched to a tax return, and in the highest-income groups the SEO sample became so sparse that matching was judged to be futile and the tax sample alone was used. Below this cut-off point, however, the SEO sample was used as the frame for the composite data base, and the matching process involved deciding, for each household, whether it should have filed a tax return, and if so, selecting an appropriate return (or returns). The SEO survey contained enough information on the amounts and kinds of income received and the demographic characteristics of household members to make an appropriate selection quite feasible. By later standards, the matching techniques used in creating the MERGE file were fairly crude. The SEO file was first converted into tax-filing units. Both samples were then divided into 74 "equivalence" classes on the basis of certain categorical variables, including whether over or under 65 years old, major source of income, marital status and number of dependants. Within each equivalence class, a tax return was then selected for each SEO unit. The primary criterion in selecting the return was taxable income; in the initial step agreement within ± 2 per cent was demanded. Among the returns within the acceptable income range, a choice was made on the basis of a consistency score, which assigned points for agreement on such attributes as home ownership and types of income reported. Using these criteria, a match was found for over 97 per cent of the SEO units on the first computer

pass. A further 2 per cent were computer-matched by relaxing the income limits somewhat, leaving one half of 1 per cent (151 cases) which were matched by hand.

17. Work on statistical matching has continued at Brookings. A second undertaking involved linking data from the 1971 CPS (containing income data for 1970) with the 1970 income tax file. Apart from updating, the new project was also designed to develop a set of generally applicable techniques that can be used to construct similar files on a recurring basis over time. It was for this reason that the files selected were chosen: both are available annually.

18. A somewhat different approach was adopted by the Office of Business Economics (OBE) of the Department of Commerce in the construction of a composite data base for use in estimating income distribution. This match involved the 1964 CPS and the 1964 Income Tax Model. The variables used were much the same as those used in the Brookings study; both procedures made use of most of the available information. But the method of selecting a match was quite different. In effect, the OBE procedure handled both alignment (under-reporting) and selection in one step by matching cases occupying the same rank in the distribution, rather than cases having the same money income. Unlike the Brookings method, which sampled the tax returns to find a match for the SEO units so that some returns were used more than once and others not at all, OBE used each return only once and every return was used. Since the sizes and weighting patterns of the two samples were different, this involved splitting individual cases to obtain the same relative weights. The results of the OBE study were published in several articles in the Survey of Current Business. Work on updating with later data is now in progress.

19. An interesting combination of exact and statistical matching techniques has been employed by SSA in connexion with a study of the impact of income taxes and social security taxes on income distribution. The construction of the data base for this study started with the 1973 Exact Match file of CPS, income tax and social security data described above. The income tax return information included in the exact-match file was limited, however, and did not include income tax liability or the desired detail on income. In order to remedy these deficiencies, detailed federal income tax return information, including income amounts and amounts of tax liability, was added by a statistical matching technique. To improve the quality of the match, however, the income tax return sample was first augmented by adding limited demographic information to it through an exact match with the SSA files using a process similar to that described above. Thus, the statistical match combined two files, each of which was the result of an exact match. The technique employed in the statistical match was a further development of the distance-function approach first employed by Brookings.

20. One final example that may be cited is the composite data base developed in connexion with the National Bureau of Economic Research project entitled The Measurement of Economic and Social Performance (MESP). The objective was to create for each of the major institutional sectors of the United States national accounts (households, establishments, governments) a general-purpose microdata base aligned to the national accounts totals for that sector. For establishments and governments (there are 75,000 governmental units in the United States), the

basic method was exact matching, using largely open (non-confidential) sources. For households, however, a statistical matching technique was developed employing somewhat different principles from any of those described above. The MESP data base is based upon the 1970 PUS, to which Tax Model information has been added. The method of matching employed rejects the distance-function approach on both theoretical grounds and on grounds of computing expense and instead sorts both the files to be merged in an order such that cases which are most similar to each other are adjacent. Determination of the variables to be used in this ordering rests on a probability analysis of the relationship between the variables which are common to both files and the other variables in each data set. After the two files are ordered on the same basis, they are merged, and the nearest cases from the two files are matched with each other. An integral part of this project has been testing of the goodness of the match. One such test employed a split-half technique. Using only a part of the data items, half the sample was matched against the other half. It was then possible to compare the imputed values of the data items not used with the actual values. It was found that in most instances differences between results obtained using the imputed values and actual values were not statistically significant.

B. The Nordic countries

21. It is expedient to consider work being done in the Nordic group of countries together, since developments in those countries have much in common. In those countries the concept of the population register is central to the statistical system. Exact matching has followed almost automatically, and there has been little interest in statistical matching. Problems of confidentiality, however, have become increasingly acute.

1. The population register

22. There is a long tradition in the Nordic countries of keeping full and complete and generally open vital records on all of the inhabitants. In Sweden, for example, parish records have been kept for centuries on births, deaths, marriages, migration and population. From 1634 on, there have been lists of the population by parish, stating name, date of birth, marital status, household composition and address for each person. In the modern period, these lists were until 1967 based on written questionnaires filled out annually by the head of each household. In years of population and housing censuses, the questionnaires were simply extended to include additional items for census purposes. Even between censuses, however, occupation and main employer were asked for. Since 1967, the lists have been computerized, and only changes were required to be reported, except in census years. From 1947 on, nine-digit identification numbers constructed from birth dates have been allocated to all registered persons, and since then most registers have been reorganized and ordered according to those numbers. In addition to the basic population register, there are various other registers containing identified individual data, including mothers of children under 16 (who receive state allowances), persons within the medical insurance system, persons on pensions, lists of assessments for real estate taxation and

assessed income for income tax. The unique characteristic of these registers is that all of them are open to the public unless explicitly prohibited by law. Access to information that may be damaging to the individual concerned is prohibited, but until 1973 exceptions were often made for recognized researchers who agreed not to reveal individual data in their research reports.

23. In Norway, a statistical file system has been under preparation in the Central Bureau of Statistics (CBS) since the early 1960s. The work on the personal data files was first concentrated on the task of establishing a complete and operational population register and promoting its use among administrative agencies. A system with unique and permanent personal identification numbers was introduced in 1964, when an identification number was assigned to each inhabitant covered by the 1960 population registration. The personal identification numbers are now used in almost all statistical surveys and censuses of persons. Even more important is the fact that the identification number system has also been adopted by many administrative agencies from which the Central Bureau of Statistics receives data. Among the most important are the population registration, the tax authorities, the national social insurance, the health administration and the school administration. As in Sweden, the data system now consists of a number of different registers relating to different topics (vital statistics, taxes, health, social welfare, crime, education), but they are related to one another through the use of the common identification number.

24. A similar system of interrelated registers also exists in Denmark. The Central Population Register includes all persons who are or have been resident in Denmark, with their date and place of birth, sex, current nationality, civil status, municipality of residence and family circumstances. Related registers include an income tax withholding register which gives details of the income of all persons over 15 years of age together with wealth status at the end of the year; a number of registers relating to education (organized by institution, by course and by student); and a register of dwellings. As in the other Nordic countries, links are maintained through the assignment of unique identification numbers - not only to persons but also to educational institutions etc.

2. Exact matching

25. The objective of the establishment of a register system like those described in paragraphs 22-24 above is of course to permit bringing together information about particular individuals from different sources. Two factors combine to make this approach feasible, where it would not be in other circumstances: use of the identification number for matching purposes is not greatly hampered by confidentiality considerations and the registers for the most part are complete enumerations. There are numerous examples in the three countries of the kinds of matches that can be and have been carried out. These include both linkages of individual records from the same source but from different time periods, and linkages of records from different sources. In some cases, they involve linking samples to registers. In Norway, the most extensive linkage of individual records from the same source but from different periods is that which has been carried out regularly since 1964 on the vital data on births, marriages, deaths and

migrations each time an updated population status file is required. The linkage of data records constructed from the 1965 Census of Fishermen with the records for the same persons from the 1960 Population Census was the first experience in linking data from two sources. Later projects have included linking the 1960 Population Census file with the population status files for the end of 1967 and later with the 1970 Population Census, as well as linkage between the income record files and the census file. In Denmark there are numerous examples of linkages involving the Labour Force Sample Survey, the Central Withholding Tax Register, local governmental registers of family allowance payments and education registers relating to students with information from the Central Population Register.

26. It is interesting to note, however, that even in countries with as advanced and open a register system as these, the ultimate step has not been taken: there are still separate registers on separate topics. The content of the basic central population register varies from country to country, but for the most part it is restricted to demographic information. Matches with other registers are made when needed for analytic use. As noted above, even for small to medium-sized countries such registers do approach the convenient limits of present computer technology, and increasing their size would increase costs without necessarily making them significantly easier to use. More important, perhaps, is the issue of confidentiality. This will be discussed in section 4 below.

3. Registers as a substitute for censuses

27. The population registers of the Nordic countries constitute a continuous record of population status. It was inevitable, therefore, that a question would arise as to whether there was any longer a need for a conventional census enumeration. This question was considered at the twenty-eighth meeting of the chief statisticians of the Nordic countries, held in 1976. In the three countries considered here, the population registers do supply the basic demographic information, but there are various other kinds of information normally collected in censuses that are less well covered. Chief among these are information on housing and some aspects of occupation and income.

28. In Sweden, some of the information required for both the 1970 and 1975 censuses was taken from registers. This included name, address, sex, age, marital status, citizenship, country of birth, some regional information and income. The most important information that was missing was on dwellings, occupation and data which connect dwelling with place of work. The possibility of replacing the census completely with register information in the longer run is under study. One requisite for this would be a register of dwellings.

29. In Denmark, although discussions of the possibility of replacing the census by a central register of persons had started earlier, a conventional census was held in 1970. By 1975, however, developments had reached the point where it was decided that no conventional census was required. A coding system for occupation and work status was under development in the central population register which, although not completed, was expected to make possible a classification of the total population by these characteristics. From 1974 on, it was in principle

possible to combine information in the occupation register with data from the register of persons by means of the information register used by the tax administration, which showed, among other things, the salaries paid by each employer to each employee. One of the major weaknesses of the register system remained the lack of a register of dwellings; the decision not to conduct the 1975 census was predicated on the assumption that such a register would be established in connexion with the cadastral survey of 1977. Since the data needs were not quite met in 1975, however, it is expected that the data requirements for the census year 1981 will be extensive. Several possibilities are under consideration. One is to supplement the register information with a broad survey to cover missing items like education and the work force. It is still possible that some information will be collected on a census basis in 1981, but if a census is conducted it will be very different from previous censuses with regard to both collection and compilation.

30. In Norway, it is not considered that the development of registers has gone far enough to make it possible to base the 1980 census on register information only. The central persons register has, however, attained a high standard and is developed with the nuclear family as the unit. Also, all completed education is registered with person identification. There are plans to develop a country-wide register system for land units, buildings and addresses through a joint reference number, connected with both the central persons register and the establishment register. In 1977 a register of dwelling construction was established for tax evaluation purposes and as a basis for municipal building registers. As a consequence of new social security rules, it is anticipated that registers of employees and employers will be established at the social security offices. It will then be possible to combine the register of employers with the CBS register of establishments so that employees can be classified according to kind of activity and location of establishment.

4. The problem of confidentiality

31. Despite the long tradition of open data, the growing world-wide concern with privacy has had a strong impact in the Nordic countries. In Sweden, the 1970 census of population and housing met with some protests, mostly from persons who objected to handing their forms to their landlords, who served as collectors. Increasingly, the widespread commercial use of registers as sources of addresses for mailing lists and by private credit investigation agencies met with disapproval. The National Central Bureau of Statistics, in anticipation of the imposition of stricter rules, began refusing access to individual data even for research purposes, although such data could often still be obtained from the originating agencies. In 1973, a data law was passed by parliament. Among other provisions, the law requires that, if they contain data on identifiable individuals and are run by means of a computer, all registers not ordered by the Government must apply for a licence to a National Data Inspection Board. No exception was made for statistical registers. Certain sensitive data are permitted in such registers only under exceptional conditions, and runs of a file against other files are prohibited in most cases. Additionally, individuals were given the right to inspect the data on themselves in any register, and this has been interpreted to include statistical registers. In Norway and Denmark, the reaction

came later and seems to have led to somewhat less stringent regulations. Statistical files have been exempted from the right of access by individuals to their own data, and in Denmark the projected law would not require the permission of the supervisory board for exact matching of data from different sources for statistical purposes. It is too soon to predict what the ultimate impact of these new data laws in the Nordic countries will be upon purely statistical uses of data, but it is clear that the issue of confidentiality will be of increasing importance.

C. Canada

1. The data base concept

32. Canadian developments, while similar in many respects to those taking place during the same period in the United States, were perhaps unique in one aspect. This was the early recognition by the then Dominion Bureau of Statistics of the data base principle and the consequent decision to commit significant resources to the reorganization of files, procedures, computer programing and modes of thinking required for its implementation. Unlike the United States experience, this was largely an internal development rather than a response to outside demands. In several early theoretical papers, the concept of a system of microdata sets organized around the decision-making units in the society (individuals, business enterprises, government) was recognized, as was the potential utility of such data for, inter alia, simulation analyses. But at least initially there was no thought of public release of data relating to individual units. Rather, the data base concept was seen as a matter of efficient methodology for the production of largely conventional tabulated output - but tabulated output which could not always be anticipated in advance. From the beginning, it was recognized that confidentiality would be a major problem, and disclosure analysis has been carried to a high level.

33. In the course of this pioneering effort, many of the problems raised in the present report, on both a conceptual and a technical level, were recognized. In some cases solutions were worked out; in others, practical compromises have been found; and some problems still remain. But the over-all impact upon the work of Statistics Canada has been far-reaching. And, as time goes on, the utility of microdata has been increasingly appreciated by users outside Statistics Canada. In consequence, Canada has become the second country to develop a public use sample of unidentified census microdata.

2. Matching experiences

34. An early Canadian example of a type of statistical match which differs from any described so far was that undertaken between the Family Expenditure Survey (FEX) and the Survey of Consumer Finances (SCF). These were both relatively small samples, one containing about 12,000 cases and the other 15,000. From the start, however, they were designed with statistical matching in view, and therefore both surveys contained information on which to base a match, including a common core of

questions on characteristics such as age, sex, occupation, labour force status, education, immigrant status and so forth. Issuing a joint questionnaire containing all the questions asked in both surveys was rejected on the ground that the response burden would have become intolerable and the number of refusals and response errors therefore unacceptable. The SCF is basically an income survey, carried out annually on a national basis. The FEX is much less frequent, having been conducted in 1970 for the first time since 1948. Both surveys relied on the same sampling frame, the Canadian Labour Force Survey. Interviews for both were carried out within two months of each other. Thus, the two surveys were conceptually aligned in the planning stage, and the sampling frame was such that they were compatible in terms of coverage. In this respect, therefore, many of the problems of consistency and alignment that other statistical matching projects have had to face did not arise. On the other hand, the choice of criteria for determining a match involved difficulties, since it had to rely on behavioural aspects of the relation of asset and debt holdings, on the one hand, and consumption patterns, on the other, to income, demographic properties and categorical variables such as home ownership. The behavioural relationships had to be established by conducting a pre-match analysis of each sample to determine the variables to be used for matching purposes and their relative importance; regression analysis was employed for this purpose. The project included an attempt to assess the quality of the matched file, and on the basis of that analysis some directions for future work were suggested.

35. A particularly ambitious exact matching project is now under way in Canada, involving the 1971 census and 1970 income tax data. A national sample from the census file of 30,000 households (about 100,000 individuals) is being matched to the complete income tax file, in a multistage process. The matching information used was name, address, sex and age, and the matching process was carried out using files containing only the identifying information. As a second stage, the statistical data from the census and the tax records is being added to the merged file. The income tax records contain a unique income tax identifier. The final stage will be to use these identifiers to add the tax records of the sample for subsequent years, 1971 to 1975, to develop longitudinal income histories. The file will thus contain socio-economic data for the individual from the census plus an income history file. One of the prospective uses of the file is to evaluate the census data on income. It is expected to be a reasonably good sample of prime-age male labour-force participants and will constitute the first longitudinal data base. Income tax records also provide data on migration and unemployment (since unemployment insurance benefits are taxable). Thus the data base has a very good analytic potential.

D. India

36. In most developing countries the sample survey is the primary source of social data. The National Sample Survey (NSS) in India is one of the earliest and most ambitious examples of a continuing multisubject survey, the first round having been conducted in 1950. In the early years, two or more rounds were sometimes conducted in one year. Since 1957, however, the survey has been conducted on an annual basis. The different rounds of the survey have covered a wide variety of

topics. Until 1974/75, some socio-economic subjects like consumer expenditure and labour force were included in all rounds, with a view to providing time series data. Evaluation of the results, however, led to the conclusion that it would be more valuable to extend the subject coverage, collecting each type of information at less frequent intervals. The present programme, therefore, envisages combining the important subjects into five major groups, each constituting a complete survey round. The first two groups would be repeated at intervals of five years and the other three at intervals of 10 years. In any 10-year period, therefore, three rounds would be available for undertaking surveys on other subjects of topical interest. The five groups are as follows:

(a) Employment, unemployment, rural labour and consumer expenditure;

(b) Self-employment in non-agricultural sectors;

(c) Population, births, deaths, disability, morbidity, fertility, maternity, child care and family planning;

(d) Debt, investment and capital formation;

(e) Land holdings and livestock enterprises.

37. The sampling frame for the survey is provided by the population census conducted every 10 years. The sample is stratified, in the first instance, into rural and urban sectors. Within each sector, there is random selection of subunits - i.e., villages in the rural sector, blocks in the urban sector - within which a final set of sample households is again randomly chosen. Sampling fractions vary from state to state, in general inversely with the population of the state. Over the years the sample size has increased continuously, from 960 villages and 406 urban blocks during 1952/53 to 8,512 villages and 4,872 urban blocks during 1974/75. The national sample is designed to provide reliable data on a national and state level. In order to provide more detail within states, however, in recent years the national sample has been augmented by a parallel survey of equal size carried out by the states, using the same schedules and procedures.

38. The survey covers, at the present time, a year-long period, usually July to June. Each sample household is visited once during this period. To spread the work uniformly over the entire year, the period of inquiry is divided into equal segments of two to three months, called subrounds. The sample units are equally distributed among the subrounds; each subround is designed to produce reliable estimates. Although each survey covers multiple subjects, different sets of households are usually selected for different subjects. It is recognized that, especially in the rural sector, household activities are interrelated, and that more satisfactory results would be obtained if information on different aspects of household activity were collected from the same sample household. In some earlier rounds, such an approach was tried, but it was given up mainly because of respondent fatigue.

39. The survey uses a moving reference period for socio-economic variables. Demographic particulars and housing conditions are collected as of the date of the survey, while data on labour force are obtained with a reference period of one week. Most of the information on consumer expenditures and household enterprises refers to a period of one month, while data on vital events are obtained with a reference period of one year. As the survey is spread throughout the year, the one-week and one-month periods of reference do not relate to any fixed point of time. This approach was, as noted, adopted primarily for effective utilization of the field force, but it is also considered appropriate to an agricultural economy with pronounced seasonal fluctuations, on the ground that a moving reference period provides better estimates of seasonal and annual values of characteristics.

40. As is to be expected over so long a period, there have been changes in the concepts of the data collected. Consumption, for instance, was in the early rounds defined to include expenditure on births, deaths, marriages and litigation, as well as house construction. These items were dropped in later rounds, on the ground that their infrequent occurence led to distortions. Consumption was, however, in all rounds recorded in detail by item, so that adjustment to comparable concepts would be possible. Changes have also occurred in the concepts of labour force and employment and in the definitions of residence and migration.

41. The primary objective of the survey exercise has been the production of pre-specified tabulations. For the early rounds, shortage of storage facilities often led to destruction of the basic data to make way for data from more recent surveys. For rounds after the nineteenth (1964/65), the primary data have been preserved and stored centrally by the National Sample Survey Organisation. In practical terms, however, access to the basic data is not feasible, and what is available for analytic use are the tabulations. The amount of cross-classification in the tabulations is naturally limited. For consumption, for instance, frequency distributions by consumption size class, on both a per capita and per household basis, are available. Since 1957/58, state estimates have been made, but there is no breakdown by any other characteristic - occupation, for instance - within states. Also, there are problems of comparability over time owing to conceptual changes; although the basic data would allow adjustment to comparable concepts, such adjustments do not appear in the tabulated output.

42. One use that has been made of the survey data has been in checking statistics derived from other sources. This has been done, for example, with regard to agricultural statistics on land utilization and crop cutting. On some topics, alternative estimates are available in the national accounts. The NSS results relating to total national consumption, for example, do diverge from those appearing in the national accounts, but the cause of the divergence is not entirely clear. During the decade of the 1950s, the two were in fairly close agreement, but since then the NSS has fallen increasingly short of the national accounts totals. The increasing divergence points up the importance of alignment and the need for a thorough study of the sources of such differences, if unbiased conclusions are to be drawn from them.

43. The NSS by now represents an invaluable body of data, over what is for a developing country a very long period of time. Though its analytic application

is restricted by the set tabulations, much of the basic data still exist so that it may be possible to exploit them further at some time in the future.

E. Brazil

44. A major effort has been invested in Brazil in a large-scale multisubject household survey, the results of which have been so organized as to constitute a continuing data resource. The national survey of household budgets and expenditure (ENDEF), involving some 55,000 households, was conducted by the Instituto Brasileiro de Geografia e Estatistica (IBGE) during 1974 and 1975. After some years of experience in household surveys directed to particular limited ends, it had gradually become clear that there were great advantages to studying households in an integrated way as producers, consumers and eventually investors, through the use of integrated hierarchical questionnaires which would make possible analyses at several levels of observation, employing a great variety of explanatory variables. It was expected that such an approach would yield information of great potential usefulness for national, regional and sectoral planning and that it could obtain types of data not well covered by more conventional inquiries. Planning for such a multidisciplinary survey necessitated many compromises, not only among the representatives of the various disciplines who wanted data on different topics but among the fundamentally different approaches of the sociologist, the statistician and the policy maker. After the survey had been completed, however, it turned out that many of the doubts expressed in advance were resolved. In the view of its sponsors, the ENDEF survey yields in fact much more than the sum of the contributions of each of the disciplines that participated in its development; it has demonstrated that a survey having a multidisciplinary approach and multiple purposes can delineate the level of life of the population and the factors explaining it and provide the information to study the explanatory factors and the corrective measures that could be developed within a humanistic view of economic growth.

45. Perhaps the most innovative aspect of the Brazilian survey, however, is not the data themselves or the method of collection but the approach taken to their preservation and use. It was recognized that the results, in terms of data use, of previous multiple-purpose surveys had for several reasons never measured up to the hopes for them. Use of ad hoc processing or special-purpose programme packages severely limited the results that could be obtained and, with the inevitable pressure of inadequate time and resources, analysts were satisfied to meet the most immediate demands. After tabulations were made, little care was taken of the archives and it became virtually impossible to rework the data. Another contributing factor had been a limited view of the possibilities of exploitation. Although multiple ends were included in such surveys, each was treated as separate in the results, without studying the interrelationships.

46. In developing the ENDEF survey, IBGE sought to rethink these problems and to try for solutions that would optimize the collection effort relative to its analytic potential - in other words, not to collect more information than could conceivably be used but at the same time to provide for future needs. In the first place, it was recognized that there is always need, in planning or even in simply

forecasting, for unforeseen tabulations and that such needs may arise years after the first publication of the results. A system that would permit continuing exploitation of the data was necessary. In the second place, to maximize the use of the data, it was necessary to provide rapid and continuing access to them without additional cost for repeating calculations and without the need for remobilizing a large administrative structure. These considerations pointed to the development of a permanent data base, but not simply a data base containing the raw data. A hierarchical structure was developed, permitting access at several levels - the person, the food-preparation grouping, the budgetary unit etc. At each of these levels, derived statistics (means, distributions, ratios) were included as well as the original observations. In addition to this hierarchical structure by level, the data base is also structured by major analytic domains, according to what are designated "themes" - nutrition, budgetary structure etc. In its final form, the data base contains about 850 observed variables and nearly double that number of computed variables, amounting to about 2 billion bits of information

F. Kenya

47. An example of a continuing survey programme much more recently established than India's is Kenya's National Integrated Sample Survey Programme. Phase 1 is expected to cover the period 1976-1981. By the end of this period, the results of the 1979 population census will make possible a revision of the National Sample.

48. The programme includes a series of sample surveys, covering a wide range of subject matter in demographic, social and economic areas. The focal point of the integrated approach is the multipurpose National Sample; the enumeration units for all surveys in the programme are chosen within this framework. The structuring of the programme ensures that interconnecting links will be established between various surveys. The use of a single multipurpose sample rather than specifically designed samples for each survey is expected to increase efficiency through more effective use of limited manpower in a number of ways, particularly in the supporting cartographic work and in more effective supervision as well as in the field operations.

49. The sample is designed to yield reliable estimates at the national level, for the urban/rural breakdown and for selected items by provinces and cropping zones. But in the first phase the sample will not be of sufficient size to provide reliable estimates at the district level. The total sample size consists of 21,000 households in rural areas (1 per cent) and 8,000 households in urban areas (2 per cent). The sample design concentrates all sampling activities within a limited number of areas, called Primary Sampling Units (PSUs). In the rural component, there are 64 PSUs. In most PSUs two Secondary Sampling Units, consisting of a cluster of about 175 contiguous households, were randomly selected. In the urban component, 80 clusters of about 100 households were selected, nearly half from Nairobi. In the larger urban areas, the clusters were stratified, prior to sample selection, by approximate economic standard; in the rural areas, by crop zones.

50. For any given survey, respondents are selected as a subsample of the basic National Sample, depending on the nature and purpose of that survey. A description of the "nesting" procedure for three specific surveys - the National Demographic Survey (NDS), the Labour Force Survey (LFS) and the Integrated Rural Survey (IRS) - may illustrate the principle. The NDS seeks to measure relatively rare events like births and deaths. It needs a relatively large sample, but the time required per interview is short since the information sought is brief and precise. All 29,000 households in the National Sample are therefore included. For the LFS, the number of respondents can be safely reduced without adverse effect upon the quality of the information obtained, but the time per interview is longer. For the IRS, the time spent per household over a 12-month period is very substantially longer still, but the number of households is much smaller. While all households in the National Sample are enumerated for the NDS, progressively smaller subsamples are selected for the LFS (10,000) and IRS (2,500). The subsample is so managed that no household is included in both of the latter two.

51. It is hoped that the National Integrated Sample Survey Programme will help to fill what is seen as a major gap in the information base for national development planning, namely the lack of pertinent and current data on socio-economic trends. Ultimately the various surveys are expected to yield information that is not only internally correct but mutually consistent - an essential prerequisite for interdisciplinary analysis.

Annex II

BIBLIOGRAPHY

Alter, Horst. Creation of a synthetic data set by linking records of the Canadian Survey of Consumer Finances with the Family Expenditure Survey, 1970. Annals of economic and social measurement, 3:2:373-394.

Althauser, Robert P. and Donald Rubin. The computerized construction of a matched sample. American journal of sociology, September 1976:325-346.

Alvey, Wendy and Cynthia Cobleigh. Exploration of differences between linked current population survey and social security earnings data for 1972. In Proceedings of the American Statistical Association, Social Statistics Section, 1975:121-128.

_____ and Fritz Scheuren, comps. Selected Bibliography on the Matching of Person Records from Different Sources. In Proceedings of the American Statistical Association, Social Statistics Section, 1974, pp. 151-154.

Armington, Catherin and Marjorie Adle. Creating the MERGE-70 file: data folding and linking. Research on microdata files based on field surveys and tax returns. Working paper I. Washington, D.C., The Brookings Institution, June 1975 (mimeo).

Bachi, R., R. Baron and G. Nathan. Methods of record-linkage and applications in Israel. Bulletin of the International Statistical Institute, XLII:2:766-784.

Budd, Edward C. The creation of a microdata file for estimating the size distribution of income. Review of income and wealth, 17:4:317-333.

_____ and Daniel B. Radner. The Bureau of Economic Analysis and current population survey size distributions: some comparisons for 1964. In The personal distribution of income and wealth, Studies in income and wealth, No. 39. James D. Smith, ed. Washington, D.C., National Bureau of Economic Research. Pp. 449-558.

_____ and _____. The OBE size distribution series: methods and tentative results for 1964. American economic review, LIX:435-449, May 1969.

_____ _____ and J. C. Hinrichs. Size distribution of family personal income: methodology and estimates for 1964. Bureau of Economic Analysis, Staff Paper No. 21. Washington, D.C., United States Department of Commerce, June 1973.

Byrne, J. Completeness of birth registration in the Commonwealth Caribbean. Paper presented at the First Conference of Commonwealth Caribbean Government Statisticians, Trinidad. Revised, 1966.

Chang, M. C., T. H. Liu and L. P. Chow. Study by matching of the demographic impact of an IUD program: a preliminary report. Milbank Memorial Fund quarterly, 47:2:137-157.

Choe, Ehn Hyun. Problems and adequacy of vital statistics in Korea. The Population Studies Center, Publication Series No. 1. Seoul, Seoul National University, April 1967.

Computer methods for family linkage of vital and health records. By J. M. Kennedy and others. Atomic Energy of Canada, Ltd., April 1965.

Coulter, Richard W. United States. Department of Agriculture, Statistical Reporting Service. An application of a theory for record linkage. Paper presented at the 4 June 1977 meeting of the Washington Statistical Society.

Exact match research using the March 1973 current population survey - initial states. Studies from Interagency Data Linkages, No. 4. By Frederick J. Scheuren and others. Washington, D.C., Office of Research and Statistics, Social Security Administration, July 1975.

Fellegi, I. P. and S. A. Goldberg. The computer and government statistics. In The role of the computer in economic and social research in Latin America. Nancy D. Ruggles, ed. Washington, D.C., National Bureau of Economic Research, 1974. Pp. 1-18.

_____ and J. L. Phillips. Statistical confidentiality: some theory and applications to data dissemination. Annals of economic and social measurement, 3:2:399-410, April 1974.

_____ and Alan B. Sunter. A theory for record linkage. Journal of the American Statistical Association, 64:1183-1210.

Francois, Patrick. Enquête nationale sur les budgets et l'alimentation des ménages (ENDEF). Working Document No. 7, Organisation for Economic Co-operation and Development, Development Centre Study Session on Multi-Purpose Household Surveys in Developing Countries, Paris, 14-18 November 1977 (mimeo).

Hambright, T. Z. Comparison of information on death certificates and matching 1960 census records: age, marital status, race, nativity and country of origin. Demography, 6:4:413-23, November, 1969.

Hansen, Morris H. The Role and Feasibility of a National Data Bank, based on Matched Records and Interviews. Report of the President's Commission on Federal Statistics (Washington) vol. 2:1-63.

Hirschberg, David, Robert Yuskavage and Fritz Scheuren. The impact on personal and family income of adjusting the current population survey for undercoverage. Paper presented at the meetings of the American Statistical Association, August 1977.

Impact of the Malaysian family planning program on births: a comparison of matched acceptor and non-acceptor birth rates. By J. T. Johnson and others. Population studies. Forthcoming.

Janson, Carl-Gunnar. Project Metropolitan: a presentation. Research report No. 1. In Project Metropolitan: a longitudinal study of a Stockholm cohort. Carl-Gunnar Janson, ed. Stockholm, Stockholm University, 1975.

Johnson, J. T., T. B. Ann and L. Corsa. Assessment of family planning programme effects on births: preliminary results obtained through direct matching of birth and programme acceptor records. Population studies, 27:1:85-96.

Kenya. Central Bureau of Statistics. The National Integrated Sample Survey Programme: sample design. Kenya statistical digest, XIV:3. September 1976.

_____ _____ Results from the first six months of the dual-records system. Demographic Working Paper No. 2., Nairobi, February 1976.

_____ _____ Social perspectives, various issues, e.g. Non-farm activities in rural Kenyan households, 2:2, June 1977; Literacy in rural Kenya, 2:3; The rural Kenyan nutrition survey, 2:4;

Madigan, F. C. The Mindanao Center for Population Studies: a Philippine POPLAB report. Scientific Series No. 8. Chapel Hill, N.C., Laboratories for Population Statistics, July 1973.

_____ and H. Bradley Wells. Report on matching procedures of a dual records system in the southern Philippines. Demography, 13:3:381-395, August 1976.

Marks, Eli S., William Seltzer and Karol J. Krotki. Population growth estimation. A handbook of vital statistics measurement. New York, The Population Council, 1974.
Specifically: pp. 101-124, 132-138 (methodology)
pp. 195-220 (matching operation)
pp. 286-298 (implementation in a hypothetical case study)

Martens, Peter. Project Metropolitan: a description of its data archive as of March 1975. Research report No. 2. In Project Metropolitan: a longitudinal study of a Stockholm cohort. Carl-Gunnar Janson, ed. Stockholm, Stockholm University, 1975.

Meeting of the chief statisticians of the Nordic countries in Reykjavik, 1976. Statistical reports of the Nordic countries, No. 32, Copenhagen, 1977.

Micro data sets, simulation and statistical systems. Paper presented at Workshop on Micro Data Sets. By T. Gigantes and others. Washington, D.C., National Bureau of Economic Research, 22-23 October 1970 (mimeo).

Mehta, D. C. Report on matching: under-registration study (pilot), urban vaso project. Report Series No. 3. Kaira Sample Registration Research Project, KSRRP (Urban)-KU(9)-2. Gujarat, Ahmedabad, Directorate of Health and Medical Services, 1967.

_____ and M. H. Shah. Report on sample registration scheme (pilot), rural Gujarat. Gujarat, Ahmedabad, Directorate of Health and Medical Services, 1966.

_____ and _____ Report on sample registration, rural Gujarat, October 1965 to September 1966. Gujarat, Ahmedabad, Directorate of Health and Medical Services, 1968.

Murty, D. V. R. and P. K. Jain. Report on pilot sample registration scheme in five villages in Mehrauli Block, South Delhi, 1 December 1963-30 November 1966. New Delhi, Central Family Planning Institute, 1967. Mimeographed report, labeled "Preliminary".

Neter, John, E. S. Maynes and R. Ramanathan. The effect of mismatching of the measurement of response errors. Journal of the American Statistical Association, 60:1005-1027.

Newcombe, Howard B. Record linking: the design of efficient systems for linking records into individual and family histories. American journal of human genetics, 19:3:335-359, May 1967.

_____ and Martha E. Smith. Changing patterns of family growth: the value of linked records as a source of data. Population studies, XXIV:2:193-203, July 1970.

_____ and James M. Kennedy. Record linkage, making maximum use of the discriminating power of identifying information. Communications of the Association for Computing Machinery, 5:11:563-566.

Nordbotten, Svein. Individual data files and their utilization in socio-demographic model building in the Norwegian Central Bureau of Statistics. Review of the International Statistical Institute, 38:2:193-201.

_____ Purposes, problems and ideas related to statistical file systems. Bulletin of the International Statistical Institute, XLII:2:733-748.

Ohlsson, I. Merging of data for statistical use. Bulletin of the International Statistical Institute, XLII:2:750-764.

Okner, Benjamin. Constructing a new data base from existing microdata sets: the 1966 merge file. Annals of economic and social measurement, 1:3:325-362, July 1972. Comments by Christopher A. Sims, Jon K. Peck, Edward C. Budd.

_____ Data matching and merging: an overview. Annals of economic and social measurement, 3:2:347-352.

/...

Ono, Mitsuo. Discussion paper on the measurement of public welfare status. Paper presented at the meetings of the American Statistical Association, August 1977 (mimeo).

Perkins, Walter M. and Charles D. Jones. Matching for census coverage checks. In Proceedings of the American Statistical Association, Social Statistics Section, 1965, pp. 122-139.

Population growth estimation. Final report of the Population Growth Estimation Experiment, 1962-1965. Dacca, Pakistan Institute of Development Economics, 1971.

Radner, Daniel B. Federal income taxes, social security taxes, and the United States distribution of income, 1972. Paper prepared for the Fifteenth General Conference of the International Association for Research in Income and Wealth, August 1977 (mimeo).

_____ and Hans J. Muller. Alternative types of record matching: costs and benefits. Paper presented at the meetings of the American Statistical Association, August 1977 (mimeo).

Rajaraman, Indira. Data sources on income distribution in Bangladesh, India Pakistan and Sri Lanka: an evaluation. Review of income and wealth, 22:3:223-238, September 1976.

Rao, V. R. and N. S. Sastry. Evolution of a total survey design - the Indian experience. Paper prepared for the International Statistical Institute, Fortieth session, Warsaw, September 1975 (mimeo).

Ruggles, Nancy and Richard Ruggles. A strategy for merging and matching microdata sets. Annals of economic and social measurement, 3:2:353-372, April 1974.

_____ _____ and E. Wolff. Merging microdata: rationale, practice and testing. Annals of economic and social measurement, 6:4, October 1977.

_____ and _____. The role of microdata in the national economic and social accounts. Review of income and wealth, 21:2, June 1975.

Salazar Carrillo, Jorge. The use of the computer in handling large price files: the experience with a benchmark collection in Latin America. In The role of the computer in economic and social research in Latin America. Nancy D. Ruggles, ed. Washington, D.C., National Bureau of Economic Research, 1974. Pp. 291-312.

Sastry, N. S. Household surveys in India: quality of data collected and their usefulness for planning and policy purposes. Working Document No. 16, Organisation for Economic Co-operation and Development, Development Centre Study Session on Multi-Purpose Household Surveys in Developing Countries, Paris, 14-18 November 1977.

Scheuren, Frederick J. and Barbara Tyler. Matched current population survey and social security data bases. Public data use, 3:7-10, July 1975.

Schiefelbein, Ernesto. A simulation model of the Mexican educational system. In The role of the computer in economic and social research in Latin America. Nancy D. Ruggles, ed. Washington D.C., National Bureau of Economic Research, 1974. Pp. 231-252.

Seltzer, W. PGE studies: cost and effectiveness. In Proceedings of the American Statistical Association, Social Statistics Section, 1971:97-105.

_____ and Arjun Adlakha. On the effect of errors in the application of the Chandrasekar-Deming technique. 1969. (Reprinted as Laboratories for Population Statistics Reprint Series No. 14. Chapel Hill, 1974.)

Shapiro, S. and J. Schachter. Methodology and summary of the 1950 birth registration test in the United States. Estadística, 10:37:688-99.

Smith, Martha E. and H. B. Newcombe. Methods for computer linkage of hospital admission-separation records into cumulative health histories. Methods of information in medicine, 14:3:118-125.

Sims, Christopher A. Comment. Annals of economic and social measurement, 3:2:395-398, April 1974.

Some Preliminary Results from the 1973 CPS-IRS-SSA Exact Match Study - Papers on the Reconciliation of Survey and Administrative Income Distribution Statistics through Data Linkage. In Proceedings of the American Statistical Association, Social Statistics Section, 1975. Especially: Scheuren, Fritz and H. Lock Oh. Fiddling around with nonmatches and mismatches.

Srinivasan, Shri K. and Shri A. Muthiah. Problems of matching births identified from two independent sources. Journal of family welfare, 14:4:13-22.

Steinberg, Joseph and Leon Pritzker. Some experiences with and reflections on data linkage in the United States. Bulletin of the International Statistical Institute, 42:786-805, 1967.

Stone, Richard. A system of social matrices. Review of income and wealth, 19:2, June 1973.

Tepping, Benjamin J. A model for optimum linkage of records. Journal of the American Statistical Association, 63:1321-1332.

Thailand. National Statistical Office. Report of the Survey of Population Change: 1964-67. Publication series E-SuR-No. 3-69. Bangkok, Office of the Prime Minister, 1969.

Unimatch 1 Users Manual - A Record Linkage System. United States Bureau of the Census, Census Use Study. Washington, D.C., March 1974.

United Kingdom. Central Statistical Office. Distribution of income statistics for the United Kingdom, 1972/73: sources and methods. Economic trends, 262:78-96, August 1975.

_____ _____ Income distribution in the United Kingdom, 1974/75. Economic trends, 282:99-101, April 1977.

_____ Department of Health and Social Security. Low incomes: evidence to the Royal Commission on the Distribution of Income and Wealth. Supplementary Benefits Commission Paper, London, Her Majesty's Stationery Office, No. 6, 1977.

United Nations. Department of Economic and Social Affairs. The Mysore Population Study. Population Studies, No. 34. 1961.

United States. Department of Agriculture. Statistical Reporting Service. Selection of a surname coding procedure for the SRS Records Linkage System. (B. T. Lynch and W. L. Arends). Paper presented at the 4/6/77 meeting of the Washington Statistical Society.

_____ Department of Commerce. National Bureau of Standards. Accessing individual records from personal data files using non-unique identifiers. NBS Special Publication 500-2. 1977.

_____ Department of the Treasury. Office of Tax Analysis. Reducing and merging microdata files. Office of Tax Analysis Paper, No. 7. October 1975.

_____ _____ _____ Statistical problems of merged data files. Office of Tax Analysis Paper, No. 6. 12 December 1975.

_____ Department of Health, Education and Welfare. National Center for Health Statistics. Comparison of the classification of place of residence on death certificates and matching census records, United States, May-August 1960. Vital and Health Statistics, Pub. No. 1,000, Series 2, No. 30. Washington, Public Health Service, January 1969.

_____ _____ _____ The Cooperative Health Statistics Program: its mission and program. Final report from the Task Force on Definitions to the Cooperative Health Statistics Advisory Committee, 30 August 1976. Department of Health, Education and Welfare Publication No. (HRA)77-1456.

_____ _____ _____ Standardized micro-data tape transcripts. Department of Health, Education and Welfare Publication No. (HRA)76-1213, February 1976.

_____ _____ _____ A study of infant mortality from linked records: method of study and registration aspects. Publication No. 1,000, Series 20, No. 7. Washington, Public Health Service, 1970.

/...

_____ _____ _____ Vital signs present at birth. Vital and health statistics. Department of Health, Education and Welfare Publication No. (HSM) 72-1043, Series 2, No. 46, February 1972.

Vincent, P. E. Une méthode préconisée en URSS pour apprécier la qualité de statistiques démographiques. In Proceedings of the World Population Conference, 1954. United Nations publication, Sales No. 55.XIII.8, vol. IV, pp. 241-47.

Watts, Harold W. Microdata: lessons from the Survey of Economic Opportunity and the graduated work incentive experiment. Annals of economic and social measurement, 1:2:183-192, April 1972.

_____ Micro-economic data banks: problems and potential. In The role of the computer in economic and social research in Latin America. Nancy D. Ruggles, ed. Washington, D.C., National Bureau of Economic Research, 1974. Pp. 57-66.

Wells, H. B. Matching studies. In Measuring the effect of family planning programmes on fertility. C. Chandrasekaran and A. I. Hermalin, eds. Dolhain, Belgium, Ordina Editions, 1975. Pp. 215-244.

_____ and B. L. Agrawal. Sample registration in India. Demography, 4:1:374-387.

Wolff, Edward N. The goodness of match. National Bureau of Economic Research Working Paper, No. 72. December 1974 (mimeo).

Zerkowski, Ralph Miguel. Administrative files and national accounts system. Paper presented at the second Latin American Conference of the International Association for Research in Income and Wealth, Rio de Janeiro, 1974 (mimeo).

_____ Possibilities and limitations of administrative files for providing social indicators. The Brazilian experience. Paper for Organisation for Economic Co-operation and Development, Development Centre Study Session on Multi-Purpose Household Surveys in Developing Countries, Paris, 14-18 November 1977 (mimeo).

Annex III

FUTURE WORK

1. The Expert Group Meeting on Methods of Integration of Social and Demographic Statistics, which was held at United Nations Headquarters in March 1978, considered a wide-ranging programme of kinds of activities that it would be useful for the United Nations Statistical Office to undertake in this area. The Group recognized that the scope of work suggested is well beyond the resources presently available and that in making an actual work plan it would be necessary to assign priorities and make choices. In some cases they indicated areas of high priority. Nevertheless, they regarded all of the areas described below as ones where an important contribution to national and international work could be made by the Statistical Office alone or working in conjunction with other organizations.

2. The main headings and the substantive content of the programme below are those agreed upon by the Expert Group. Within the main heads, an attempt has been made to organize the work into logical units. The four main groupings are (1) substantive subject-matter questions, (2) organizational and procedural problems of data collection and management, (3) statistical methodology, and (4) computer technology.

1. Substantive subject-matter questions

3. The proposals in this area inevitably overlap other aspects of the ongoing work of the Statistical Office. What are touched upon here are those aspects which are particularly important in connexion with integration. Three areas of work may be identified.

(a) Concepts and classifications for specific social fields

4. A high priority was attached to the ongoing work on the classifications and concepts for specific social fields, considered in terms of both internal needs and cross-field needs, including data inputs and kinds of outputs for typical uses. This work should include such matters as the definitions of units like the family and the household and the treatment of changes in such units over time. A progress report on related work in this area is before the Commission (E/CN.3/518) and work on this project is expected to continue.

(b) The structure of linkages among data bases

5. The objective of work in this area is to show how classifications and reference and reporting units in different areas can be interrelated so as to permit movement from one field to another while still retaining flexibility within fields, and how alignment of data from different sources can be accomplished while retaining the integrity of the original sources. The role of the national accounts as an organizing frame for social microdata and in relating micro- and

macrodata and the potentialities of further subsectoring of the national accounts should be explored. The question of time, including not only time budget studies but time in a longer perspective (as in lifetime income) and time as an organizing principle, should be considered. To a large extent this topic is a further development of the present document, and it might suitably be the subject of a report for a future session of the Commission.

(c) Analytic uses

6. The work programme proposed at the first meeting of the Working Group on the System of Social and Demographic Statistics of the Conference of European Statisticians in 1973 had included a review of the main classes of analytic models employing social and demographic data, and the Expert Group considered that it would be appropriate to pursue this avenue now. In addition, it was thought that the Statistical Office might usefully broaden its co-ordinating role by promoting a dialogue between producers and users of data. Discussion of the role of statistics is needed on the international level, it was thought, and for this planners and statisticians must be brought together and statisticians must get into analysis. Efforts should be devoted to promoting analytical work by statisticians directed towards foreseeing the needs of users.

2. Organizational and procedural problems of data collection and management

7. The kinds of problems envisaged here are those of data-base management in an organizational and procedural sense, not a computer technology sense. They would include methods of achieving standardization and consistency, documentation requirements, access, disclosure control and confidentiality protection. In each of the areas listed below, it would be useful to start with a survey of country practices. On the basis of such surveys, it would then be useful to issue technical reports. These are not areas where international guidelines are appropriate, but a discussion of possible alternative procedures may be very helpful to countries.

(a) Documentation and archiving

8. A study in this area should cover data description and documentation (both human-readable and machine-readable) required for permanent archiving and for working files. It should discuss the functions of the data administrator and the data librarian and should consider problems of achieving consistent procedures that have arisen in practice.

(b) Confidentiality and privacy

9. Confidentiality as a policy issue and its implications for statistical legislation should be considered, with a review of how these problems have been dealt with in various countries. Techniques of disclosure control should be described. The issue of privacy - intrusiveness - as distinguished from that of confidentiality or disclosure should be discussed.

(c) Uses of administrative data

10. This topic should deal with practical questions of obtaining access to administrative data and influencing the development of administrative records in ways that are suitable for statistical use. It should review country practices to discover what techniques have been successful in specific instances, what have failed and what can be suggested as guidance to countries for the future.

3. Statistical methodology

11. This heading is meant to include only those aspects of statistical methodology that are particularly involved in the construction of microdata bases and their integration with aggregated data. The problems involved are ones where technical studies would be appropriate. Most of them would require the services of consultants to provide expertise not available among the Statistical Office staff.

(a) Techniques of estimation

12. This topic was considered very important. It should include techniques of editing and imputation of missing values, exact and statistical matching, problems of alignment and reconciliation of data from different sources and problems of statistical quality control.

(b) Logical file structure

13. Included here are the logical, conceptual aspects of constructing microdata bases so as to permit maximum flexibility of access and use at various levels of aggregation.

(c) Problems arising in the use of administrative data

14. Effective use of administrative data requires the solution of a number of statistical methodology problems, such as conversion from one type of unit to another, consistency of coverage etc.

4. Computer technology

15. This is an area where a great deal of work is going on elsewhere, and it is not intended that such work should be duplicated. What is included here are those aspects of computer technology that are peculiarly related to the production, storage and dissemination of social statistics and, in particular, the preparation of manuals and technical reports which will make the results of such work available to social statisticians in developing countries.

(a) Survey of work done in developed countries

16. It was considered important that work done in developed countries on machine-readable data-base construction and management (from a technical computer-oriented point of view) be brought together and colated, to make it accessible to statisticians working in developing countries.

(b) Compendium of technical information on development of machine-readable data files

17. In addition to the survey of country practices, it was thought that a compendium of technical information applicable to machine-readable data-base construction written in non-technical language would be very helpful.

(c) Techniques for various stages of development

18. This study should devote attention to computer techiques that are appropriate for countries at various levels of statistical development. Examples of machine-readable data files and data bases and their uses should be developed. Resource requirements in various practical contexts should be studied.
