PROBLEMS IN THE ESTABLISHMENT OF NATIONAL DATA SYSTEMS

Anders S. Lunde, National Center for Health Statistics

A number of countries have established extensive national data systems which include the complete registration of the population and the accumulation of comprehensive population data files through censuses, surveys, and the recording of social and economic events affecting the person. By record linkage processes these data are being combined to provide new statistical information primarily for administrative and statistical purposes but also for research. In order to bring these systems to their present state of sophistication, a number of previously unrecognized problems had to be overcome. More problems remain; some are not fully understood while others are political rather than statistical in nature.

National data systems have been most successful in developed countries with relatively homogeneous and small populations, but the concept of a total national data system has now been accepted not only by large advanced nations but by less developed countries. The principal problem for developed countries is that of recognizing the limits of any national statistical system. In less developed countries the main problem is the lack of basic data subsystems upon which the more sophisticated systems are built. For all countries, numerous methodological problems exist. The further development of national systems for some will depend upon the outcome of the issue of confidentiality; for others, on the improvement of presently incomplete component data systems.

Definition of National Statistical Data System

The national statistical data systems referred to in this paper are those which identify each individual in the population by a unique number provided each person at birth or at in-migration to the population, and use this number for linking together records from data sub-systems: - vital statistics systems, population registers, censuses, and surveys, and other data sources such as health, welfare, education, and administrative records. It is the personal identifying number which characterizes those national data systems and which at the same time creates a problem because not only is it a technological key to computer files but it also seems to be a kind of access key to the personal lives of individuals in those files. There is general agreement, however, that a number best serves as an identifier in data base computer systems. It is easier to handle for purposes of comparison and sorting, storage capacity is expanded, and records matching and linkage potentials are considerably improved. 1/

Dr. Lunde is Acting Director of the Office of International Statistics, National Center for Health Statistics. The opinions expressed are those of the author and do not necessarily reflect the views of the National Center for Health Statistics, and the Department of Health, Education, and Welfare.

The number as a computer file key in data systems must have certain characteristics which are specifically delimiting as there can be no exceptions to the definitions. The number must be unique and independent, that is, it is the sole identifier for one particular individual and for him alone. It is permanently assigned; while it is true that several numbers could be serially identified and selected on data files, each additional item or operation adds to possible error. It must be reliable in the sense that the system provides for such quality controls that the number appears appropriately on the named person's records. It is a universal number in the sense that it is one identifiable unit in a universe of numbers, such as in a definable population of persons.

The number is assigned in various ways. (See Table I) In some countries (Norway, Sweden, Denmark) the number is assigned at birth and contains within it a code showing the date of birth. 2/ In Norway, a male child born on August 26, 1975, would be indicated by the number 26.08.75 003 29

Day	Month	Year	Serial No.	Check Digits
26	08	75	003	29

The serial number (003) is sex-coded with even numbers for women and odd numbers for men. The extra digits are for a computer check in what is known as a Modulus 11 system. In other countries (Israel) an arbitrary number which has no personally identifying codes is issued at birth and to arriving immigrants. 3/ All countries make allowances for additions of this sort, including older persons born before the establishment of the system, and also for the retirement of numbers in the event of emigration or death.

The unique personal number is placed on all pertinent documents in the national data system. Theoretically, this could include an extensive list of documents numbering in the billions in some countries but generally they are used on readily identifiable records belonging to already established administrative and statistical subsystems. Census and survey schedules; certificates of birth, death, marriage, and divorce in vital statistics registration systems; and records from such nationally important systems as the military, judicial, electoral, taxation, social security, and educational systems, are among the first to be included in national data systems. The stated primary purpose of establishing integrated systems is the improvement of national administration, evaluation and planning. Statistical analysis and research use are secondary considerations. It should be noted that the development of national systems involving the numbering of individuals is usually accompanied by the introduction of national numbering systems for industries and business establishments, and

considerable interest has been shown in creating special manpower, occupation and industry studies through the linking of both personal and industrial data files. In all countries numbers are affixed to personal and commercial documents within the context of laws and regulations which usually contain restrictions to limit abuse of the system.

Another element is the fact that the number is the key in the computer linking of records. As it is theoretically possible to apply the number to many documents pertaining to an individual, it is also theoretically possible to link a large number of separate individual records which exist in various collections or sub-systems. Some countries have combined vital statistics and census data; others have tried to clear social security files through reference to mortality files; a few have begun to link health records from various sources. 4/ There are all kinds of possibilities in such linkage for administrative and research purposes and the end is not yet in sight. Perhaps the most dramatic view of future potential is that a perfect national data system could not only serve as a continuous population register but as an "instant" population, housing, industrial and agricultural census.

Despite the considerable interest these systems, particularly the Scandinavian, have produced around the world, and the enthusiasm with which some countries have embraced the numbering system concept, many problems are associated with the creation and continued development of these systems. It is the purpose of this paper briefly to examine some of these problems in established and emerging national data systems.

National Data Systems in Operation

Perhaps the earliest continuous national system was that of France, established in 1941 for social security purposes, and based on an I.D. Number of 15 digits. A complex code for geographic areas, including department, overseas territory, city, town, and commune of birth, created many matching problems. Computerizing the system in 1971-72 was an expensive and time-consuming task. One drawback in the system is that the files at present are not cleared by death; consequently, there are more persons in the population file than actually live in France or its possessions.

In developing its advanced system in 1964, Norway found that the date of birth, the core of the unique number, was incorrect for some 100,000 persons (representing about 3 percent of its total population). This caused trouble in the longitudinal linking of data; also, all changes in the central files had to be communicated to the users. 5/ The Central Bureau of Statistics also reported that there was lack of correspondence between the statistical registers (census data, vital statistics, and population registers) and the administrative registers (education data, military and manpower data, health, occupation and industry data). It was discovered that linking these records was no easy task; in addition, the quality of the data from some sources was in doubt. One official raised the problem of the timeliness of the data and subsequently of any register and raised the question of the huge cost of maintaining up-to-date national data systems. 6/

Other countries have experienced problems in other areas. Israel reported errors in basic records and in transcribing, coding, and card-punching, all of which led to matching errors. Sweden's first unique number was inadequate for matching purposes because no computer check digits were included; this has subsequently been corrected.

When early errors have been resolved, problems continue to emerge especially when systems are being expanded, which generally means that other sub-systems are to be absorbed and the records linked. Linkage refers to the accumulation and combination of records in which matches can be made; matching is the correspondence of distinct items in different records pertaining to the same person. In any accumulation or merging, records sometimes are linked that do not match and there is a failure to link records that match. At times different definitions apply in different systems and the same conventions in recording information are not followed. There is always the problem of compatability, and of determining the denominator for the computation of rates. In many countries the first enthusiasm for the unlimited expansion of national data systems has changed to the sobering realization that the inclusion of new data sets will mean extensive adjustment and considerable work. At a certain point, the benefits from expanding the system do not justify the required expenditures. Statistical operations are everywhere limited by national budget priorities.

As a matter of fact, T. F. Hughes has recently raised questions regarding the meaning of "integration," and the automatic linking of all subsystems and integrating them into one operation for national statistics. He feels that this concept of "process integration" has been abandoned because the complete integration of the statistical system is impractical, too difficult, and probably unnecessary. He suggests that a certain amount of integration is desirable but the problem centers around the amount of integration which should be applied. Before the advantages of any comprehensive system may be enjoyed, Hughes warns that many problems of statistical organization and management must be overcome, and that the following improvements are required in statistical services: better coordination of data; better communications and and more consultations with providers of statistical data and with users of statistical information; more effective control of privacy protection and the maintenance of confidentiality; more effective management of the statistical computing service; and more involvement of statisticians with computing. 7/

National Data Systems Being Planned

A number of new problems have been identified in those countries which, lacking the technical, processing, and procedural capabilities of more advanced countries, have nevertheless moved rapidly into the computer world. With particular reference to Latin America the problems range from those associated with operations management and records administration to technical work at all levels and computer operations.

A major problem lies in the fact that few developing countries have any statistical data systems that can function as sub-systems in a comprehensive national system, with the possible exception of a national census. In Brazil a government office has completed studies to reorganize the national statistical offices into one coordinated record-linking system. The trouble is that the quality of the statistics from some agencies is seriously in question; vital statistics data, for example, are quite incomplete and inaccurate. The important statistical problems of compatability of data and data matching and linkage have yet to be faced. It is in comparison with the western European countries with centuries of record handling experience behind them that the problems of the less developed countries stand out in glaring contrast. The lack of fully operating national data component systems is a major limitation to further development.

Argentina in 1970 established a personal numbering system under the Registro Nacional de las Personas to establish a continuous population register and provide a basis for national administration and planning. Every person is issued a number at birth and the entire population is expected to be registered by 1980. The system at present functions primarily as an identification system and its full potential in social and economic planning remains for the future. Officials admit that one problem is the overwhelming records administration problem which involves handling the great flow of documents from registration offices to the central files. It has become almost impossible for this otherwise highly efficient office to implement the more comprehensive program.

Lack of trained manpower is admittedly a crucial problem and as a result, a number of data systems have not become fully functional. The manpower lack is felt at all levels, from executive and supervisory statisticians to statistical clerks, and from computer programmers to computer operators. Technical training programs which could be stimulated by United States expertise unfortunately are blocked by the language barrier. The program to arrange for the "Closer Communication, Interchange, and Joint Development of Statistics and the Statistical Profession in Latin America and the United States" sponsored by the American Statistical Association is one of the most important ventures of the association, if it will help build bridges from this country to Latin America. The Mexican Pilot visit of last May, which the Mexicans called the First International Symposium of Statistics, resulted in the organization of the first Mexican Statistical Association and in the stimulation of statistical interest not only among the professionals in government and industry but among the educators and students of the discipline. The extension of these interchanges to include Argentina, Brazil, Chile, Colombia and Venezuela next year will no doubt have similar effects.

Jumping into the computer ocean without the lifejacket of adequate planning and preparation has created a number of problems not unfamiliar to many of us in the United States. Less developed countries have purchased computer hardware far in advance of what is required and lack the manpower to use it. As a result, computer companies have been accused of over-selling their products. At a meeting of the Pan American Health Organization Regional Advisory Committee on Health Statistics, held in Washington, D.C. in January 1975, on the subject of statistical information systems, the recommendations included a series of warnings regarding the acquisition of computers and endorsed programs which would provide Latin American countries with information on the limitations as well as the capabilities of computers.

Mention has been made of the problem of complex identifiers. One of the most recent nations to develop a unique personal number for its citizens, Uruguay, has announced the creation of a number that will identify individuals through codes referring to the Spanish name system. Six digits will designate date of birth. Another six will distinguish persons as native born Uruguayans, foreigners, and naturalized citizens, and a seventh will identify sex. Four letters will follow indicating given names, father's surname, and mother's maiden name (by first letter of name). An I.D. card will be issued to all persons over 15 years of age and at first the number will be used for administrative and identification purposes. It will be interesting to see if the introduction of this administrative system, and similar systems in Latin America, will eventually develop into statistical systems which will produce data for demographic analysis and national planning as in those countries where integrated national statistical systems currently are in operation.

Political and Policy Problems

Policy problems relate primarily to specific data access and use areas and not to the establishment of national data systems as such. For the most part, the overriding considerations are the protection of the privacy of the individual and the confidentiality of the information contained in the record.

The issues begin to pile up with the question of a unique identifier; the idea that a person may be provided a single personal number (despite all its advantages considering the proliferation of credit card and other personal numbers) creates a problem in some countries, whereas in

countries where citizens are accustomed to identity numbers, this is not an issue. There is some confusion in the United States regarding the use of a unique or standard identifier. On one hand, the Secretary of the U.S. Department of Health, Education, and Welfare in 1973 supported the report of the Advisory Committee on Automated Personal Data Systems, which recommended that a system using a standard universal identifier should not be established in the United States. 8/ On the other hand, the Federal government itself has extended the use of the Social Security Number (SSN) legally and extralegally. Some subnational groups use this number; for example, the automobile drivers licenses in the District of Columbia contain the SSN as the license number and file number. The SSN, however, in its present form, does not meet the criteria for use in a national data system. It is not unique; millions of persons have two or more numbers. It is not universal in coverage. It could be modified for use in a national data system; although such modification would be expensive.

Another numbering system exists in the United States, but is has not been nationalized or legalized in Federal statutes. Since the 1940's, States have been interested in a universal birth number for identification of persons and record linkage. Since 1967, by agreement of State Registrars, an 11-digit file number appears on birth certificates, which includes codes for State of birth, year of birth, and serial number. Interest in the numbers waned following World War Two but has risen again. The numbering issue will re-emerge in connection with discussions on the statistical requirements for a national health insurance system. For the first time, aside from the census, all persons living in the United States, regardless of age, may have to be registered to receive health benefits. Such registration will have to be continuous as newborns move into the system and deaths are subtracted from it. Massive records search and retrieval mechanisms will be established and record linkage will be a basic necessity. It is inconceivable that the creation of a unique identifier for computer use will not receive serious consideration.

In every data collecting country, in varying degrees, two problems have emerged: the invasion of privacy problem and the confidentiality problem. The first represents a conflict between the government's "need to know" and the individual's right to determine what personal information he will provide. The second refers to the degree to which replies in statistical data collection can be kept confidential, that is, will not be used to the detriment of the individual but for statistical purposes only. Statistical purpose in this case means that the data will be aggregated, or summarized, for public use without reference to individuals. It is not intended to elaborate on these issues; they are now being discussed around the world. As regards statistical techniques and other means of handling the problem of the invasion of privacy, Dalenius has recently provided a useful overview. 9/ On the issue of confidentiality, Martin has discussed the role of statistical legislation in safeguarding confidentiality and encouraging trust in official statistical organizations. 10/

Many countries are involved in these debates. What is perhaps the most sophisticated statistical data system in the world has been planned in detail by the Federal Republic of Germany; it is as yet nonoperational because the law to establish the system has been strenuously debated in Parliament for over two years, the principal issue being confidentiality. Similar debates have taken place in Belgium, the Netherlands, and Luxembourg. These Benelux countries have planned an international data linking system; but the Netherlands, remembering the invasion of World War Two, has been concerned about the security of the records and the possibility of their unauthorized use. In the United States, both issues, in terms of statistical data collection and data use, have been widely discussed throughout government and in Congress. The latest outcome of this public activity is the Privacy Act of 1974 (P.L.93-579), by which Congress seeks to safeguard privacy from the misuse of Federal records, to provide individuals with access to records, and to establish a Privacy Protection Study Commission. Government statistical agencies have created special committees to study the issues. The American Statistical Association has established an ad hoc Committee on Privacy and Confidentiality to review and evaluate the statistical implications of recent acts and reports and current proposals in the Office of Management and Budget.

With the extension of national statistical data systems (as well as their establishment) by the merging of data from various sub-systems, these issues will continue to rise. It is incumbent on statisticians to explain the need for the data, justify the use to which it will be put, establish its confidentiality through statistical legislation and, as Dalenius puts it, take a public relations approach "directed toward the benefits of statistical production rather than the potential dangers." 11/

Concluding Remarks

1. The focus on problems of the national statistical data systems in this paper should not detract from the fact that the systems are operative in many countries and are providing data for administrative, statistical, and research purposes. In some cases, new insight into social and economic activity and their relation to family formation, fertility, employment, income, and a host of other factors, are made possible. Sweden, for example, produces a miniature data archive, a population register based on all the persons born on the 15th of any month. The register is updated every year with data from a variety of sources; censuses, vital statistics, income statistics, and so on, and serves as a sampling frame for demographic and economic studies. 12/ And as regards problems, they are generally contended with, and overcome.

2. Almost all extensive national data systems, especially those referred to in this paper, have been made possible by the use of a standard identifier in the form of a unique number for each person. It would appear that a numbering system is required for the efficient operation of a sophisticated national data system in which records are to be linked.

A very complex number, which would seem to emphasize uniqueness, may instead create many problems for quality control. When numbers contain digits coded to individual characteristics, they may be less efficient than an arbitrarily assigned number. Any error in the birth date, century, geographic code, sex, ethnic group, family name, etc. means a change in the number and in all corresponding files. Some designs date from earlier technologies: -sound systems, punch cards, and the like; modern computer technology does not require any backup if the number is unique. Therefore, although there may be special reasons for designing a number containing various codes, it seems that a simple non-coded number is to be preferred. Computer check digits should, of course, be included. This problem of the best characteristics of a number; necessary digits, number codes, etc., has yet to be resolved, and may be a matter to be decided separately in terms of each country's requirements for items of identification.

3. The principal problems in national systems remain statistical in nature and are not unknown to statisticians. At the basic level they relate to problems of data system design and operations, forms design, data collection, coding, data processing, quality control, and so on. On another level they deal with problems of linking of records and matching of items, with data compatability, timeliness, and problems related to the merging of statistical file systems.

4. One problems remains for those countries without integrated systems. Do all countries have to develop comprehensive data systems as are being developed, for example, in the Scandinavian countries? Hansen has stated the view that in the United States the additional statistical output that would result from the linking of federal records would be less than anticipated, and that more can be achieved by linking particular records when they are needed. 13/ It will be important to observe how countries like the Federal Republic of Germany (62 million), with populations larger than Norway (4 million), Sweden (8 million), and Denmark (5 million), handle the complex problems of data file combinations in newly established integrated systems.

5. Problems related to invasion of privacy and confidentiality will crop up increasingly throughout the world. These questions will be raised wherever established systems are under investigation, where new systems are being established, or old ones expanded. They are also stimulated by a political climate of suspicion and uncertainty regarding the intent of data collection by the government. To allay public distrust it will be necessary to establish confidence as regards individual rights and the security of records which provide information on persons. An emphasis on use of information only for statistical purposes must be made apparent.

6. There is much to learn from those countries with integrated national data systems in terms of the design of the system, the role of the sub-systems, the integration of data, the control of data, statistical techniques, computer techniques, and techniques related to the combination or merger of various data files from different sources. As other countries extend their collection and analysis mechanisms to include more social and economic activity of the population, whether they develop integrated or segregated data systems, they will learn much from the experience of those nations which have pioneered the development.

****

References

1. Lunde, Anders S. (1975). "National Data Systems and Record Linkage". Bulletin of the International Statistical Institute, Warsaw (to be published).
2. Aurbakken, Erik and Petter Jacob Bjerve (1973). "The Role of Registers and the Linking of Data from Different Sources". Bulletin of the International Statistical Institute, Vol. XLV, Book 3, pp.169-184.
3. Ohlsson, Ingvar (1967). "Merging of Data for Statistical Use". Bulletin of the International Statistical Institute, Vol. XLII, Book 2, pp. 750-764.
4. Bacci, R., R. Baron and G. Nathan (1967). "Methods of Record Linkage and Applications in Israel". Bulletin of the International Statistical Institute, Vol. XLII, Book 2, pp. 766-785.
5. Lunde, Anders S. (1975). "Birth Number Concept and Record Linkage". Journal of the American Public Health Association, August (in press).
6. Aurbakken, Erik and Petter Jacob Bjerve (1973), op. cit., p. 172.
7. Nordbotten, Svein (1967). "Purposes, Problems and Ideas Related to Statistical File Systems". Bulletin of the International Statistical Institute, Vol. XLII, Book 2, pp. 733-748.
8. Hughes, T.F. (1975). "Possible Implications of Integrated Systems for National Statistical Services". Bulletin of the International Statistical Institute, Invited Paper No. 48 (in press).
9. U.S. Department of Health, Education, and Welfare (1973). Records, Computers, and the Rights of Citizens. Washington, D.C.: DHEW Publication No. (OS) 73-94, July 1973.
10. Delenius, Tore (1974). The Invasion of Privacy Problem and Statistics Production--An Overview. Stockholm: Statistisk Tioskrift 1974: 3, pp. 213-225.

11. Martin, Margaret E. (1974). "Statistical Legislation and Confidentiality Issues". International Statistical Review, Vol. 42, Number 3, December, pp. 265-281.

12. Dalenius, Tore (1974), op. cit., p.233.

13. Ohlsson, Ingvar (1967), op.cit., p. 754-55.

14. Hansen, Morris H. (1971). The Role and Feasibility of a National Data Bank, Based on Matched Records and Alternatives, in Federal Statistics: Report of the President's Commission. Washington, D.C.: Vol. II, pp. 5-61.

TABLE I. ITEMS IN IDENTIFICATION NUMBERS IN NATIONAL DATA SYSTEMS

Country	Total Digits	Date of Birth			Serial Number	Control Digit	Sex	Place of Birth
		Day	Month	Year				
Europe								
Denmark	10	2	2	2	3	1		
Finland	10	2	2	2	3	1		
France	15		2	2	3	2	1	5
Federal Republic of Germany	12	2	2	2	4	1	1	
Iceland	9	2	2	2	2	1		
Norway	11	2	2	2	3	2		
Sweden	10	2	2	2	3	1		
United Kingdom	8[a/]			2	3			3
South America								
Argentina	8[b/]			(2)	8		(1)	
Brazil	11				9	2		
Chile	8[c/]				7	1		
Uruguay	13[d/]	2	2	2	(4 letters)	2	1	
Asia								
Israel	6[e/]				6			
Japan	14[f/]	2	2	2	3	1		4

Other Countries Planning or Discussing Systems

Belgium	Switzerland
Netherlands	Spain
Luxembourg	South Korea
German Democratic Republic	United States
Portugal	

a/ Five letters and 3 numbers (LMNOP 123) including codes for sub-district of birth, year of birth registration, and quarter of registration.

b/ For cross-checking purposes, has the capacity to add 2 digits for year of birth and 1 digit for sex (1=male, 2=female).

c/ New numbering system will include codes for: place of birth, Civil Registration Office where birth was filed, and year of registration.

d/ Nine numbers and 4 letters; first 6 digits are date of birth, 7th is sex and nationality, followed by first letters of four names, and 2 check digits.

e/ Has no code for identifying characteristics.

f/ One of the numbers reported being tested in urban areas.