Demography India
Vol. 21, No. I (1992), pp. 65-71

C. S. Arora*

Use of New Technologies for Future Indian Censuses—The Possibilities and Challenges

Introduction

JUST about a century ago, it was a census employee from US Census Bureau, Herman Hollerith, who had set the ball rolling as a pioneer of new techniques in information processing, when he developed a method of recording census responses by punching holes in response cards and feeding them through an electronic tabulating device to capture the data. Hollerith later founded a company that eventually became IBM. However, perhaps even he did not anticipate that a century later a person residing in any part of the world and interested in contacting another person in Thailand would be able to do so through a unique Population Identification Number stored on mainframe computers with on-line storage capacity of 100 giga bytes. The Central Population Data Base maintained by Thai Government in fact, contains information relating to possession of fire-arms, facial description and finger prints, driving licence particulars, insurance and immigration data also besides the usual demographic parameters. Such has been the influence of technology on the population related data processing.

Technological innovations have been the cornerstone of the census processing particularly in the last three decades or so. The general introduction of large scale Batch-oriented data processing methods on computers around 1960s was perhaps one large important step forward towards the automation for census data processing. The range of automation and technological options for various stages of census data processing from keyboard data entry to online computer coding and electronic dissemination is so diverse and baffling that the challenge now lies in the proper choice and management of the technology rather than the bits and bytes of the technology itself. In fact, in a tendency for copying the technologies followed in most developed countries, sometimes the "me too" syndrome has led to a situation where rather than "managing the technology", some of the situations have led to the "technology managing us". Herein lies the challenge to understand the evolving technologies and to use these technologies to advantage in our situations.

This paper outlines various possible new technologies for the important stages of census data processing for future Censuses alongwith various issues involved. While outlining the

* C S. Arora, Director EDP. The views expressed in this paper are from the author in his individual capacity and do not necessarily reflect those of the Office of the Registrar General, India.

advantages, the present limitations and the factors to be viewed for caution are also attempted to be described. Finally, the importance of the "non-technology factors" is also stressed towards the choice of the new technologies for future Indian Census in the 21st century keeping in view the fact that the chosen technologies have to address towards the test of processing over a Billion records of Population Data.

Enumeration Block Mapping

The first potential stage for automation and application of newer technologies is for generation of Enumeration Block maps, which are used to allow enumerators to reach their areas and through which it is made sure that the whole territory is covered without any overlaps or omissions. Such Enumeration Block maps require of course, the updation by the cartographers from one census to another. If done by conventional manual method, the whole process is tedious involving safe keeping of hard copy maps over a number of years and the need for redrawing and updating everytime because of deterioration of readability and the boundary changes. Modern computer methods developed for Geographical Information Systems can allow maps to be drawn on the digitizers or scanners and stored on computer media. Besides providing the maps on a good quality of paper, updating the data becomes very easy and the storage can be compact and well organised. Creation of automated mapping system can be one of the most dramatic technological innovations in the future censuses to develop computerised maps including all important boundaries and land-marks sucfi as roads, rivers, lakes etc. This one technological advance in itself will save considerable time, increase accuracy and reduce monotonous manual work content.

Coding Methods

Though hypothetically ideal, completely pre-coded census schedules are not possible atleast for Indian census which collects a range of diverse data on economic questions, languages and migration etc. Verbal answers have to be collected in the field for these items to be converted into codes for further computer processing. The process of converting the verbal answers into codes is being done manually so far which of course, is a laborious, error-prone and monotonous job. Technological advance for this stage of processing may either be in the form of "computer assisted" or "automatic coding". In computer assisted coding, it is still the coder who determines the code to represent a particular verbal description of an item. Automatic coding will, however, not require any human intervention and hence these two methods are fundamentally different in their operation.

In computer-assisted coding, the coder may key in the whole or the essential part of the verbal description written on the schedule. The computer will then suggest one or several codes together with their verbal equivalents from the computer maintained code directories. This may allow the coder to point out the appropriate choice. If the coded text is insufficiently specific, the system may suggest that the coder provide ancillary information. Obviously, this requires ancillary information also to be available while working on the computer system. Here it may facilitate if the data items not requiring coding arc already stored in the computer so that the system may consult these also for supplementary information. This has

an implication that the data entry of those items which do not require coding takes place separately in advance or the task of keyboard data entry and computer-assisted coding are combined. However, since this form of coding requires considerable judgement skills on behalf of the coder, it may not be efficient to use him/her for simple keyboard data entry. This has a managerial/organisational implication. Regardless of the high skill level and subject knowledge of the coder, computer assisted coding will still lead to a number of cases where the coder is not able to assign a code. The process anticipates flagging such cases for later resolution by a subject matter expert and resultant rework.

In automatic coding, the software system theoretically has access to all information which a human coder would use like code directories, neighbouring items in the schedule, information pertaining to several household members and even households and persons of similar characteristics. A computer system decides automatically which code to assign. Though attempted by some countries like Switzerland and Canada using the recent technical developments in the areas of Artificial Intelligence and Expert Systems, it may still be a far cry for India to go in for automatic coding atleast for 2001 census.

The technology of computer reading of handwriting has not yet been sufficiently resolved to the stage where computer can directly understand the schedules collected from the field and code them automatically.

Data Capture

Inspite of the dramatic and rapid advancements in Information Technology, providing inputs to the computer system continues to be a bottleneck. For 2001 census population of India heading towards a billion, data inputting for computer processing is the single most important stage requiring consideration not only at technical level but also on organisational, managerial and financial levels too. The manual key stroking is a tedious and slow operation and a technological breakthrough has to be applied at this stage towards bringing out the tabulations faster. For handling the workload of a billion records of data entry, the challenge is centred around two major questions - "where" and "how" would it be done? The "where" issue involves the degree of centralisation or decentralisation by way of number and locations of the data entry and processing centres. The "how" issue involves the technology to be used for converting the data from the coded schedules into computer readable files. In this direction, Optical Mark Reading (OMR) process for direct inputting of data into computer seems to offer more flexibility in decentralising the operation. However, some of the concerns about the cost, environmental control requirements, paper quality and printing accuracy etc. have to be resolved first by way of a pilot field trial for '^proving" the technology before a decision for its full scale field adoption in the 21st century census. The problems that may occur in using this technology particularly may sound to be pedestrian in nature but they can cause serious disruptions in census processing. Seemingly mundane problems like insufficient cleaning of printing presses causing traces of wrong ink to appear as marks, improper transport or storage of schedules, inadequate coding of schedules, wilful damage to sheets by operators etc., could defeat the very purpose of adoption of the technology. To properly understand the experiences of other countries and to come out with

deliberated choice based on initial field trials, a project has now been envisaged in the office of the Registrar General, India in this regard.

Optical Character Recognition (OCR) and Optical Mark & Image Reading (OMIR) are the other variations of the data entry technological break-throughs already tried to various degrees of success by countries like Hongkong and Japan. These variations of technology are also proposed to be duly examined in detail for their possible application for data inputting for the Indian census in 2001. The OCR technology is of course, an important pre-requisite for automatic coding referred earlier and it requires writing the alphabets in a pre-structured way for being acceptable to the machine.

Voice Recognition Technology is another technological advance for the data entry capture stage. However, keeping in view the current status, these systems which are still in experimental stages have limited vocabulary of isolated specific words and so atleast for 2001 census in India, full scale implementation of this technology is not envisaged atleast at this point of time. There are, however, promising developments which may produce a breakthrough in near future.

Data Editing and Correction

Though the uniformity of concepts and maintenance of certain standards of accuracy are achieved through manual checking and supervision prior to computer inputting, the computer based data editing of the census data is considered a very vital and important stage towards bringing out the final tabulations. In the computer edit programme, all those errors which could develop in the data in the intermediate stages of processing are also visualised alongwith the possible method of imputations. The computer editing would aim to correct the errors if any, in these data fields so as to appear more logical. Since the data fields in a record would be very much inter-related with one another, errors in them could be detected through inter-consistency checks and all such errors could be controlled to a great extent through preliminary and final edit of the data on computer. Needless to say, the knowledge of the subject matter expert goes heavily in the computer edit programme and its resolution during processing. Keeping in view the inherent nature of editing and the imputations involved, the computer editing and correction has to be "on-line" inter-active process, though theoretically, in completely automated system, batch processing is also possible. Going by the present software engineering trends, the data editing and correction programmes would be more user- friendly, portable and memory efficient in the future censuses. In fact, a significant advance has already been made for 1991 census data editing itself by proposing to use CONCOR Module of IMPS software of micro-computers. There is of course, considerable scope of improvement for bringing the on-line transaction to be more user-friendly.

Data Storage and Management

Optical disks offer potential for vast amounts of storage at a relatively low cost. Already today's optical disks can store 200 million bytes of data or enough characters for about 1000 books. However, today's optical disks are "read-only" and at the utmost, one can write data

once. Optical disks that can be erased are still in the laboratories and might be available for use in future censuses. This storage technology will probably eliminate or drastically reduce the use of magnetic tapes in census data processing and will eventually replace magnetic diskettes though probably not until the onset of 21st century. The optical disks offer many exciting possibilities for the census organisation for the storage and distribution of gigantic Data Bases. In Indian context, though we may not rule out the adoption of optical disks for the main processing computers, the continuation of high-density conventional magnetic storage will be continued atleast partly. Optical disks are however definitely envisaged for data dissemination as described below.

Data Dissemination

By far, the prevailing mode of dissemination by most of the census organisations in the world has been the printed publications. The obvious disadvantages are, of course, the long delays because of volume involved and the associated paper and printing costs. Magnetic tapes as the dissemination medium are free from these drawbacks but are rather suitable only for large users who have the computer processing equipment.

The widespread availability of micro-computers has now made it possible for the data users to receive the basic files on floppy diskettes for their further analysis and special tabulations. Advances in optical storage technologies particularly CD-ROM, which can roughly store equivalent of 1500 floppies have now offered powerful technological alternative to the data users for their further analysis. For giving the basic micro-data files to the users in the readily accessible form, appropriate control and accounting procedures within the confidentiality constraints have however, to be inbuilt in the software for creation of the new data files. It is envisaged that in future censuses, a Data User Services set-up headed by a subject matter specialist in the Office of the Registrar General, India would play a leading role for dissemination of the data to the user community and this set-up would also be equipped with adequate technical facilities for generation of various data products.

Population Database

Providing on-line access to the census data through electronic transmission is another technological and powerful alternative for the users to get Census data for their further processing. This of course, requires reliable data communication lines and by design, the on-line services should be based on the equipment separate from the main computer systems used for census data processing to eliminate all possibilities of outsiders accessing confidential files or for creating other disruptions. In Indian context the satellite based computer communication network, NICNET will be able to link various points in the country upto the block level before 2001 census and the data bases on demographic indicators could be stored in the network for access by the potential users "on-dial basis". Other networks like I-NBT being introduced by the Department of Telecommunication based on packet switching concept might also act as vehicles for data transmission. The data users could also get connected to the users of Remote Area Business Message Network of the Department of Telecommunications for sharing their data bases, including the population data.

Maintaining a major on-line service is however, a big task. The users take the services provided for granted. However, they will complain bitterly when information is delayed or the system breaks down temporarily halting the services. These considerations alongwith the degree and extent of usage and the technical capability available for manning and operations of the system will however, decide the actual shape of things. In fact, the contemporary technology of Public Videotex! systems for providing a simple on-line service to the general public has already been introduced in India and by 2001, enough maturity would be achieved in videotex! area also for the creation of population data bases. The Population Identification Number Project (PIN) referred earlier relating to Thai Government's efforts in creation of the Central Population Data Bases has already achieved a milestone in the technological advance with the project being announced as the winner of the 1990 Computerworld Smithsonian Award in June 1990. The Thai project has been recognised as the world's first population database comprising 55 million people and 10 million households and a major step towards the transformation of information technology for national planning. The US $31 million project aims at creating a fully integrated demographic database consisting of 6 major subsystems: a central population database of Thailand, a personal identification database, a surname database, a marriage/divorce database, an eligible voter database and a fire-arm database.

Serving as a beacon, this project has led to contemporary efforts in this direction by various other countries. Though requiring tremendous inter-ministerial coordination within Government and outside with political parties, creation of a population data base for future censuses of India can definitely be a reality subject to the availability of the resources and the policy decisions. The technology will not be the constraint.

Management Information Systems for Census Administration

The 1991 Census employed over 15 lakh persons including 12 lakh enumerators involved in the job of visiting houses and filling the schedules for every individual from the babe in the cradle to the oldest citizen. Overseeing the mammoth job of conducting census can definitely be easier if some techniques of information technology are employed for planning, monitoring and administering the exercise. Future Indian censuses can probably rely on an elaborate automated management information system to see that the key dates in the census are met. In addition, in the planning for the future census, suitable MIS systems can perhaps also give the cost and progress data on real time basis. Evolving such systems require two specific bodies of knowledge: detailed insight into the conduct of the census operations and the skills for developing the suitable project management software on computers and their operation. In none of these two bodies of knowledge is India lacking for their application in their future censuses. A clue or two can however surely be taken from advanced countries like U.S.A. who have already demonstrated the experience of such techniques in their Census.

The "Non-Technology" Factors

Technological advances are rapid and accelerating and they provide the fountain-head for future challenges; but the integration of individual technological developments into the

appropriate working system is the major management task. As mentioned earlier, our task is to understand the new technologies and to use it to our advantage. The challenge is to manage technology rather than letting it manage us. The installation of sophisticated equipment in an environment with a work force unfamiliar with the concepts and the techniques would defeat the purpose. Hence, the challenge lies in ensuring the choice of the proper technology for each stage of census processing suitable for the organisation by establishing the strategic vision and appropriate assessment techniques. It is well known that the most powerful system in the world is as effective as its users and so the deciding factor for gaining spectacular success is not in the technology *per se* but the wisdom and the practical insight with which it is applied.

Judicious selection of particular technologies and their effective employment in a particular stage of census processing requires a careful review and consideration of a number of historical, organisational and managerial issues as well. Last but not the least, it is the financial resources and the human resource capability which determine the degree of effective technological advancement or otherwise. Against this backdrop, the Indian census organisation has been gearing up to induct the new technologies selectively for various stages, albeit, in a cautious way within the available resources.

Conclusion

While the census processing is not a- new activity, a range of new technologies and 'methods is becoming available that have the potential to remove some of the existing bottlenecks in the census processing process. In the adoption of the new technologies for various stages, however, it has to be kept in mind that the census occurs only once every ten years. It is a massive undertaking and there is little margin for failure. Hence the choice of the technologies for the future censuses must be based on minimum risk, proven equipment and adequate back-ups.