UNITED NATIONS SECRETARIAT
ESA/STAT/AC.163

Department of Economic and Social Affairs **October 2008**
Statistics Division **English only**

2010 World Population and Housing Census Programme

Report of the UNSD-UNESCAP Regional Workshop on Census Data Processing: Contemporary Technologies for Data Capture, Methodology and Practice of Data Editing, documentation and archiving

Bangkok, Thailand, 15-19 September 2008

Table of Contents

INTRODUCTION

Objectives of the Workshop

1. The purpose of the Workshop was to present international standards for processing population and housing censuses and to highlight the significant additional capabilities of contemporary technologies and their use for census data capture and data editing. More specifically, the Workshop: (1) covered the revised international standards for conducting population and housing censuses, focusing on recommended core topics as identified in the *United Nations Principles and Recommendations* Revision 2; (2) Discussed contemporary technologies in census data capture, including the use of Optical Mark Recognition (OMR), Optical Character Recognition/Intelligent Character Recognition (OCR/ICR), Internet data collection, use of handheld devices for data collection, (3) Discussed the process stages for data capture, (4) Presented an overview of major commercial suppliers for data capture; (5) Elaborated the principles and practices for census data coding and data editing; and (6) Presented international standards for data documentation and archiving and corresponding tools.

Attendance

2. The seminar was attended by 45 participants from 18 countries, by three representatives from international /regional organizations (UNFPA, UNESCAP and UNSD), and by 5 representatives of commercial providers (DRS, Betasystems, Kodak, ReadSoft and Top Image). For complete list see of participants see Annex.

Opening

3. Mr. **Pietro Gennari**, Director of the Statistics Division of the UNESCAP, welcomed the participants to the Workshop. He explained that this workshop was indeed a great forum to share knowledge and experience, identify best practices, and help strengthen national technical capacity to apply appropriate information technology effectively during the 2010 round of population and housing censuses. He also welcomed the representatives from the private companies that were specialized in producing some of the most advanced data capture technology and underlined the opportunity to learn from them about the development in this special field of technology, and the promises that these technologies offer for quality data collection, processing and dissemination, especially through the upcoming population and housing censuses.

4. He reminded that the potential of many of the IT tools – from GIS to data capturing, editing, documenting and archiving technology – is enormous for improving our capacity to produce quality and timely statistics. However, while some national statistical systems have benefited from these technology development in recent years, many others, particularly those in low-income countries or countries with economies in transition, have lagged far behind. This workshop provided an opportunity for countries to work together to assist these countries in their efforts to catch up with some of these new technologies.

5. He emphasized that this workshop was not only very important but also extremely timely, as most of the countries in the region were still in the process of preparing for the planned censuses that are to take place over the next few years. Many of them were in urgent need for technical support. The 2010 Round of Population and Housing Censuses offer a unique opportunity to further strengthen national statistical capacity and to improve the statistical basis for monitoring social, economic and environmental

development, especially progress towards the internationally agreed development goals, including the Millennium Development Goals.

6. He explained that ESCAP was committed to supporting the 2010 World Population and Housing Census Programme in the Asia-Pacific region. Since 2004, efforts have been made to identify country needs for support with conducting the upcoming population and housing censuses. ESCAP has conducted three Expert Group Meetings – in 2004, 2006 and 2007 respectively – to consult with countries on the priorities and specific activities for a regional census programme. He noted also that ESCAP conducted two information surveys – one in 2005 and another last year – on countries' census experiences and plans, with a particular focus on identifying national census-undertaking capacity and needs for support so as to facilitate knowledge-sharing and country to country cooperation. Throughout these EGMs and information surveys, countries repeatedly stressed the importance of IT applications for census management and operation, requested technical support and expressed the desire of learning about best practices in different areas. This explains why in the proposed ESCAP regional census programme is included "facilitating the effective use of information technology" as a prominent component.

7. Mr. Gennari reminded that this workshop was the second collaborative endeavour within less than a year between the United Nations Statistics Division and the ESCAP Statistics Division in the area of IT application for population censuses. The first joint workshop tool place in October last year, which focused on GIS-based data collection, analysis and dissemination. He expressed his pleasure for this collaborative work and that ESCAP could make substantive contributions to this process, in promoting microdata documentation and dissemination in this region as part of the global Accelerated Data Programme.

8. Mr. **Jean-Michel Durr**, on behalf of Dr. Paul Cheung, Director of the United Nations Statistics Division, welcome the participant and thanked them for participating in this workshop and sharing their experience with their colleagues from other countries.

9. He reminded the participants that this workshop was part of the 2010 World Programme for Population and Housing Censuses, as initiated by the United Nations Statistical Commission in March 2005 for the period 2005 to 2014. As part of this Programme on censuses, the United Nations Statistics Division conducted a series of regional workshops over the last two years, having had the two themes of the Principles and Recommendations for Population and Housing Censuses (2006) and Geographic Information Systems and Digital Mapping (2007), respectively. For the year 2008, the theme of the set of regional workshops was to present international standards for processing population and housing censuses and to highlight the significant additional capabilities of contemporary technologies and their use for census data capture and data editing. The logic was to follow the census process and to address countries' needs in their preparation of the next census. In that regard, many countries had expressed the need to take into account the technological advances made since the previous round, especially in the area of data capture and processing and requested UNSD to prepare specific guidelines, including best practices and the strategies for evaluation of different contemporary practices.

10. He stressed that data capture was one the most critical activities of a population and housing census. Capturing the huge amount of information collected in a census to convert it into a format that can be interpreted by a computer was a critical phase of census, costly and time consuming. Rapid advances in data-capture technology,

especially optical, had greatly increased the speed and reliability of producing census databases in an accurate and timely manner. Nevertheless, many countries faced in the recent past difficulties in mastering these technologies, sometimes by lack of preparation or insufficient knowledge to avoid the numerous pitfalls.

11. He presented the agenda of the workshop and explained that the workshop would alternate presentations, intervention of experts, country presentations and will also allow a session for commercial presentations.

12. He highlighted the collaboration between the UNSD and the UNESCAP for the preparation of this workshop and expressed his appreciation to the colleagues of ESCAP for hosting the meeting and providing facilities.

PRESENTATIONS AND DISCUSSIONS FROM THE VARIOUS SESSIONS

Session 2: 2010 World Population and Housing Census Programme - Preparation of the 2010 round of censuses in the region

13. An overview of the World Population and Housing Census Programme for 2010 was presented by UNSD. The three essential goals set for the 2010 programme were reiterated and the specific role of the UNSD in respect of these was outlined. UNSD has recently published the second revision of the Principles and Recommendations for Population and Housing Censuses and released it this year. UNSD, in partnership with the UNICEF and UNFPA, are developing dissemination software called CENSUSINFO, based on the original DEVINFO, but with some improved functionalities considered more appropriate for census data.

14. A round table meeting, which was moderated by UNSD, was organized to allow participating country representatives to shed light on their preparatory activities for the 2010 round of censuses.

15. **Afghanistan**: the census was initially planned for 2008 and has been postponed to September 2010, or six months after elections to ensure smooth conduct of the census free from any political interference. The following tasks still remain: update of the household listing, update of village lists, update of EA and CA Maps and offices set up: Regional Offices for the 8 regions and Districts Census Offices (398 districts). International Consultants will be recruited, an expert in Data Processing and an advisor in capacity Building. Strategies for security compromised areas have to be prepared. A transport plan is under development as well, for example for Air and Ground Transport Contracts for Deliveries of Census Materials. The Media Communication Packages are under development. Procurement plan and Training plan are also under preparation. The estimated budget for the entire 4 year program is approximately US $ 69 million.

16. The difficulties and challenges lie in security and financial issues. About 39.97% of the districts have security problem. Consultation with governors and other local authorities that was done in 34 provinces 10 provinces revealed security problems of varying degrees. In terms of budget, there was a shortfall of funds in 2008. The activities that will be funded are the following : recruitment and payments of functionaries, printing of census materials, setting up of District Census Offices, transport of census materials, media outreach, data entry processing and analysis, publication and dissemination of results. Among other difficulties is the fact that the voter's registration has been scheduled for 2008. Registration is a political process and may affect the security of people who will be carrying out census activities as well as the cooperation of respondents. The methodology is traditional: enumerators will visit every household in

every settlement in Afghanistan. The enumeration will include the nomadic population of Afghanistan, the population in institutional living quarters and diplomatic staff assigned outside Afghanistan. Manual data entry will be used (CSPro) and appropriate data processing software for tabulations and data analysis. The census in Afghanistan will cover a population over 28 million, about 5 million households, more than 21,000 EAs situated across a geographical spread of 34 provinces, 364 Districts, 34 provincial capitals, and over 42,000 villages.

17. **Bangladesh:** The exact date for the next census has not yet been decided. It might be January – March, 2011. The census is planned on a six years plan: two and half years for preparation and three and half years for data editing, capturing, processing, analysis and preparation of report. The budget is around 1237 million Taka (more than 18 million US$). The main difficulties are expected in densely populated cities like Dhaka and Chittagong. Other issues can be non-cooperation from the respondents, budget constraint, and unavailability of educated enumerators. A 100% count by *de facto* method will be adopted in the population & housing census. Zero hours of the night of the census day will be considered as the census moment. On the other hand, the hours between census moment and the census day morning (5.00 a.m.) i.e. morning of census day will reckon as the census night. To avoid omission and duplication, people will count where they will be found in the census night. Moving and homeless people will be counted in census night at transit stations like railway stations, bus stations, launch ghats, air terminals, mosques, temples, hat-bazars, and footpaths. The technology used will be GIS for production of digital EA maps, Supervisor maps, Union/Ward maps, Mauza/Mohalla maps and OMR/ICR technology is expected to be used for data capturing. In terms of dissemination: preliminary reports, national and local reports will be available both in CD and hard copy. Analytical Reports will be published on the Web.

18. **Bhutan** conducted last census in 2005 and will carry out the next one in 2015. The budget of the 2005 census was provided half by the government and half by UNFPA. GIS and GPS technologies were used in the last census for the census cartography. Data capture was manual using CSPro for data entry and SPSS for tabulation. Hundred percent data verification was carried out for the 2005 Population and Housing Census of Bhutan. GPS technology was used to map each and every structures / buildings in Bhutan during the census.

19. **Brunei Darussalam** plans to schedule its next decennial population census tentatively, in the year 2011. There is also alternative consideration to hold the Census in 2010. The Department of Statistics is in the process of preparing a project document which include work plan for activities of Census 2010 round to be submit to the management for discussion; proposed of budget etc. Issues and problems encountered in the previous Census will be addressed to the National Statistics Coordinating and the National Census Committee. There is a need for improving strategies for effective work plan, strengthening capacity building in all aspects, including ICT, reduce time lag of data release, explore new technologies of census data capture processing, and improve dissemination A Post Enumeration Survey (PES) will be conducted after the census. The 2010 round Census questionnaires is expected to be a longer questionnaire, more lengthy than the previous one, with new additional questions on ICT, disability, international migration and other users' questions. For the next Census, GIS Mapping will be used.

20. The 2010 round Census will follow the same manual data capture by personal interview. In addition, self enumeration by internet will be proposed for computer

21. **Cambodia** conducted its census in March 2008, on a *de facto* methodology, 10 years after the previous one. The total cost of the census was 7 million USD, supported by UNFPA, JICA, Government of Japan, German government and Royal government of Cambodia. The main difficulties were encountered in remote areas, with not enough infrastructure. The large scale of operation entails too many steps of training. The mapping was undertaken using GIS technology (ArcGIS 9.0) and data capture, editing and imputation manually using CSPro 3.3. The first results were released in September.

22. **China:** the next census is planned for Nov. 1, 2010. The project is under review of the State Council. Enumeration will use a short form and long form, and a death population form. Foreigners living in China will be enumerated. OCR data capture will be the main data entry method.

23. **India** is planning to conduct its next census on 2nd to 28th February 2011. With more than a billion people, covering a vast number of diverse languages, the population census is one of the world largest administrative and statistical exercises in the world. The 2001 census employed 2 million enumerators and this number is likely to increase in 2011 census. The census is conducted in two phases: first house listing operations and second enumeration of the population. The canvasser method is adopted to collect information from each household. The census represents the largest administrative exercise in the country and is included under the Union or Central List for governance. The Central Government conceptualizes the operation, finalizes the questions, supervises and provides resources for conducting census. State governments provide entire manpower for satisfactory execution of the operation. The House listing operations are conducted 7 to 8 months in advance of census taking and help to prepare the frame for collection of information during next phase of enumeration. It consists in listing each and every building, house and other structures in the Enumeration Area and collects information on their use. Population Enumeration uses the extended de-facto method. The enumeration period lasts three-week, followed by 5 days of revision round to update the population on the reference date. The current status of the 2011 Census is as follows: house listing operations will be undertaken from April to September 2010 and Population Enumeration from 1st to 28th February 2011. The revision round will take place from 1st to 5th March and the reference date will be 00:00 hrs of 1st

March 2011. Preparations begun in 2007: pre-testing of selected questions undertaken, consultation with data users held and questions proposed have been finalized after considering the feedbacks. Choice for data capture technology has been decided after examining the innovations in the field since 2001 Census.

24. Some initiatives for the 2011 census in India: During 2011 Census, information would be collected on persons present in the household at the time of enumeration including those who are visitors and have usual residence elsewhere. In case of visitors, the respondents will be asked to give address of the place of usual residence so as to enable reallocation of these persons enumerated as visitors, to their places of usual residence. Usual residents will be defined as continuously living for more than 6 months. Another new initiative is to prepare Ward maps in major towns in digital format for better coverage and enumeration. Geo-referenced EA maps will be provided to enumerators for use in census. Other initiatives include use of bar codes on census schedules for better management of inventory is contemplated and outsourcing of non-key census activities (like, printing, delivery and collection of Schedules and on-line MIS reporting). Preparation of 'National Population Register' along with 2011 Census is contemplated, which could lead to register based census in future. A proactive data dissemination strategy will be proposed as well as archiving of information in electronic format would also be given priority.

25. **Indonesia** : BPS – Statistics Indonesia is now preparing the Census of Population 2010. The 2010 census day is expected to be some time in May or June 2010 (tentative). Preparatory efforts started in the mid of 2007. Topics to be covered have been discussed including the review of lesson learned from previous Census of Population 2000. Project team has also been formed. Its main task is to conduct necessary preparation including network planning, census methodology, coverage of census variables, design of questionnaire, evaluation of enumeration area map alternatives, study of data processing alternatives, publicity and socialization strategy, census pilot study, capacity building, and budget planning. The projected number of population in 2010 is around 236 million with around 64 million of households spread over 33 provinces, 456 districts, 5,900 sub-districts, 72,000 villages, and 660,000 to 900,000 enumeration areas. The census is projected to spend about 600 billion US dollars.

26. The 2010 Population Census of Indonesia is planned to consist of two kinds of enumeration coverage, namely (i) complete enumeration to households and household members using a questionnaire with limited number of questions; (ii) sample enumeration to selected households which are planned to be 10% of total households using a questionnaire with more detail questions. Some activities related to the 2010 Population Census have been started such as data collection called "Potensi Desa 2008" which translation is "Village Potency 2008". This activity is aimed to collect information from all villages in Indonesia about village infrastructure that can be used for some purposes as a basis for "Urban/Rural" determination. The other progressing activity is updating enumeration area maps including updating maps of provinces, districts, sub-districts, villages, and some smaller administrative areas. The development of Census Data Processing Design has also been in progress. The development team is expected to do a comprehensive study of data processing alternatives, although data capture technique using scanner

and OMR, ICR technology would be preferred. Lessons learned from previous Population Census 2000 indicated that the enumerator’s handwriting quality is unfavorable to the use of ICR, hence the use of OMR might be a better choice. However, the Census team expects that respondent names are required to capture through ICR. The data processing for Population Census 2010 is planned to be decentralized into eight processing centers spread over main islands in Indonesia. Part of the reason is that it would be costly to invest one or more scanners in each of 33 provinces in Indonesia.

27. In **Iran**, there have been 6 rounds of population and housing censuses: in 1956, 1966, 1976, 1986, 1996, and 2006. Recently according to a presidential decree which was aimed at updating the information on Iranian population, the National Population and Housing Census will be conducted every 5 years in Iran and consequently the next Population & Housing Census will be conducted in 2011. Studies for the last census were started in 2004. Due to planning of the 2006 National Population and Housing Census, a specialized planning committee, was established including 4 work groups, namely geographical information, data processing, information, dissemination and training. A sample survey was associated to the Census for collecting more detailed information on some specific items. The general activities for conducting the Census were as follows: Studies for implementing mixed census and sample survey, research for selecting and using various statistical data collection technologies, studies for using CAC technology in data coding were undertaken first. A pilot census was carried out in some provinces in 2005. The strong points of the 2006 National Population and Housing Census were an appropriate design of the questionnaire forms, a good practice in offering an acceptable estimation of scanning and reading times, acceptable estimation for required staffs, improvement in quality of the developed software based on feedbacks and experiences gained from the pilot census and experimental versions, a completely automated management systems to distribute the tasks, suitable educational aids. The problems encountered were a certain lack of experience in ICR technology, some unexpected software and hardware errors, and changes in the ICR software due to modifications in the questionnaire format. (In fact it was the first experience of the SCI in application of ICR technology and some minor modifications were necessary). Due to time constraints, experiences from the Pilot Census and resulted solutions were not tested in a separate pilot census before the main Census to determine the degree of their efficiency. The budget allocated for the 2006 National Population and Housing Census in Iran was about USD 39 million.

28. .In **Korea** (Republic of), the census is conducted every 5 years since 1925. The next census (18th Census) will be conducted in 2010, November 1. The administrative organization of the census is the following: Population Census Division of KNSO, 16 Cities & Provinces, 234 Local Governments, 3,573 branch offices of local governments. The Census covers all areas incorporated within the scope of the administrative jurisdiction of the Republic of Korea. The enumeration units are: population, households, and housing units. Foreigners are also counted, while diplomats are excluded. The budget is estimated at approximately 200 million US $. Korea National Statistical Office will conduct the 2010 population census using the traditional approach. After the 2015 census, it could be changed the methodological approach from the traditional approach to a register-based approach. Various

methods will be used for data collection: mail, Internet, self enumeration, face-to face (interview). Among the difficulties expected, there are criticisms on the huge cost of census, and it is becoming increasingly difficult to interview respondents because of the increases in one person households, especially elderly people, and because of the growing awareness of privacy. The technology for data capture is not decided yet. Dissemination will comprise publication of printed tables and reports, CD-ROM, On-line dissemination.

29. **Malaysia** will conduct the next census in 2010. It will cover every person who resides in the country with a *de jure* approach, where all persons are enumerated according to their usual place of residence. The place of usual residence is defined as the place where the person lives 6 months or more in year 2010. The estimated timetable is: household listing and mapping activities, from January 2007 to December 2009 – first time implementing integrated listing (listing of establishment + household), the census day will be July 1 of 2010, field operations will last 2 months (August and September), data will processed from October to December and the preliminary report is expected for January 2011. The final report will be published in August 2011. 84% of the budget will be absorbed by wages and salaries, 6% for rentals, 3.5% for assets, and the rest for other expenses. The main issues are the budget reduction, and the innovations contemplated. Regarding data collection, it is planned, in urban areas, to use self-enumeration with drop-off and pick-up and also to propose enumeration by internet. For data capture, the office intends to use ICR.

30. During 20th century, **Mongolia** organized population censuses 9 times. The last census was organized in 2000 collecting very comprehensive population information. The reference timing for the census is non transferable night time. That means that new births after reference time and deaths before reference time will not be counted. The reference time for 2010 census will be 10 Jan 2010 at 00:00. The census operation will start at 8 a.m. of 10 January 2010 and continue for 7 days until 17 January 2010. This is the time the country has the least movement among population. The pilot census will be conducted in January 2009 and it will cover 1.5 percent of total population. GIS will be used to come up with the detailed census maps and to determine location of population and households. The new advanced technology will be used in data collection, processing, and dissemination. The Census Bureau will organize systematic dessimination of census results soon after the census results are available. The following dessimination tools will be used: publication of census monographs, CD-ROM and DVD containing census results, and also special census reports on the demand of the users of census results. In addition, the census results will be dessiminated through census website. The national and local level seminars and workshops will be organized among the users of census results. The main census results will be processed at the national and aimag level and monographs will be prepared and published for the national and aimag level census results. The monographs will be published in Mongolian and English. DEVINFO (CensusInfo) application will be used to dessiminate the census results.

31. The main features of 2010 census in Mongolia are: it is the first census to implement a new census law; a master plan will be prepared for the first time; the census will be guided by new UN recommendations. The questionnaire will include

questions on disabled people; a new communication strategy will be used to enable public participation in census activities; GIS will be used to come up with the detailed census maps; the Census Bureau will utilize advanced technologies in data processing, exchange information between national and local census bureaus, which would enable: (i) the preliminary census results to be processed very quickly; (ii) improve coverage of the census; (iii) reduce the cost for transportation and communication. The Census Bureau will closely coordinate with the Mongolian Embassies and consulate offices in order to fully cover mongolian nationals living abroad in census enumeration. It is estimated that 13.2 billion MNT (approximately 11 million USD) will be required to organize the 2010 Population and Housing Census. About 27.6 percent of the grand total or 3.6 billion MNT (approximately 3 million USD) will be from the national budget. NSO will collaborate with international and bilateral donors to seek their support in allocating 72.4 percent or 9.6 MNT of the total census budget. Among the difficulties faced are the lack of required equipments and software, insufficient skilled staff to work on GIS and lack of finance.

32. The last census in **Myanmar** was conducted in 1983. There is no plan for the next census yet, but the statistical office is considering methodologies and technology capabilities that could be used.

33. **Nepal** conducted its first population count in 1911 and since then every ten year. The next census is planned for 2011 and the preparation is already underway. The estimated budget for 2011 census is about 20 million US $. This budget covers the expenses of introducing new technology (like OMR) and infrastructure development (office building) as well as human resource development.

34. **Philippines** conducted the census in August 2007. It was originally planned for 2005. The population count was proclaimed in April 2008 at 88,570,000. Data processing was decentralized at provincial level (80 processing stations sites), regional level (17 data processing centers) completed by 2 central data processing centers. Coding is manual for place of school, highest grade, occupation and place of work. Data entry was manual using CSPro.

35. In **Sri Lanka**, the census will conducted in 2011 by the Department of Census and Statistics under Ministry of Finance. The date for next Census is not yet decided, probably in June or July. A steering committee has been appointed as well as a team for the mapping process. Mapping work has already been completed in three districts and complete mapping work will be completed in July 2010. The enumeration will be conducted under the *de jure* methodology. Data capture will use Optical Data Capturing method and more precisely Intelligent Character Recognition. A decision was made to outsource all parts of the Optical Data Capturing process. It is planned to decentralized Optical Data Capturing process at province level. There are nine provinces in Sri Lanka. The data editing process will use Census Survey Processing System (CSPro 3.3). A team is to be appointed to develop Automated Data Coding System (Occupation & Industry). The budget is estimated at 10 million US$ (1000 million SL Rupees).

36. The next Population and Housing Census in **Thailand** will be conducted in 2010, on July 1st. The preparation extends over 2008 and 2009, including a pilot census. Data collection and processing will be conducted in 2010 and in 2011 reports and

results will be disseminated. The methodology follows two steps conducted at the same time: listing stage and enumeration stage. The first one uses the Listing form for counting the number of population and dwelling units, and the second the Enumeration form for collecting basic and more detailed characteristics of both population and housing. Enumeration will be multimodal: face to face interview and self-enumeration (mailed or dropped-off questionnaire, telephone and internet). Two pilot tests will be conducted: one in 2008 in the Phitsanulok province and the second in 2009 in Bangkok. The objective of the pilot tests is to test all stages (preparation and planning, data collection, data processing and data analysis and dissemination). Data capture will use the ICR technology. The main problems encountered in the 2000 census were: undercount in the Bangkok Metropolis and other big cities, undercount of international migrants, high turn over rate of enumerators and field staff.

37. In **Vietnam** the next population and housing census of Vietnam will be conducted in 2009, with a reference data of April 1, 2009 0 hour. In order to improve the effectiveness of the census design and reduce costs, it is intended to apply the long form/short form approach for questionnaire design. A comprehensive questionnaire (short form) contains only some core questions to interview the whole population. Apart from core questions as in the short form, sample survey questionnaire (long form) with sample size about 15 percent will cover the questions on marital status, qualification, employment, fertility, death, housing. This approach used in many countries in the world, but was applied only in the last two census of 1989 and 1999 in Vietnam. But there were only two questions on fertility and deaths included in the sample with sample size of 5% in the 1989 and 3 % in the 1999.

38. It is expected, from a selection of appropriate technology of data processing and full utilization of the GSO’s informatics facilities, to speed up the release of census data. Moreover, GSO considers that the Intelligent Character Recognition (ICR) technology will be used for data capture. The use of ICR might require a huge amount of money for purchasing the facilities such as appropriate scanners, servers, client-PC, software. The United Nation Population Fund (UNFPA) will support the conduct of the 2009 population census to collect adequate and precise data on age, gender, and other relevant indicators such as ethnic minorities, migrants, male/female ratio at birth, etc. Support will also include piloting new methods for data collection, processing and analysis.

39. Highlights of the discussions and issues raised as a result of the country statements were summarized in the conclusions and recommendations.

Sessions 3, 5, 11, and 17: Country presentations and experiences on data processing

40. UNESCAP presented a summary of the results of the Status of Data Capture Technology in Population and Housing Censuses in the ESCAP region, which was sent to the countries prior to the Workshop. Forty NSOs replied. In the next census round, 11 countries will use manual data capture (18 Countries used it in the last census); 19 countries will be using OMR/OCR/ICR data capture technologies in the next census round. Two others may also. Three of the 19 (Malaysia, Mongolia and Sri Lanka) did not use automated capture in the last census. With respect to OMR & OCR/ICR the data capture process is often done in-house with some countries outsourcing. Further to the technologies chosen, 11 countries will use a

questionnaire posted on the internet as well; 2 countries will use administrative records; 5 countries will use Computer Assisted Personal Interviews (CAPI); 2 countries will use Computer Assisted Telephone Interviews (CATI).

41. **Afghanistan:** The presentation by Afghanistan focused on data processing activities for the household listing, pilot census results, population and housing Census and National Risk and Vulnerability Assessment (NRVA) Survey. It outlined the activities of the Central Statistics Office Data Processing Center data capture for household listing data pilot census data, population and housing data and NRVA survey data. The CSO Data Processing Center was built to facilitate Census and Surveys data processing and an overview of its activities, organizational structure and infrastructure setup was provided with explanation on software used and data processing methods. This was followed by an overview of the workflow, issues and constraints of 2007 pilot census. Manual data entry is intended for the 2010 census and the presentation outlined the data processing workflow for the 2010 round. The presentation concluded with an in-depth overview of a specific scanning method used in the NRVA and discussed detailed characteristics of the method such as design, read, scan station and verification.

42. **Bangladesh:** The presentation by Bangladesh outlined the historical use of OMR and OCR technology in their census and other statistical exercises. In 1981 the Bangladesh Bureau of Statistics (BBS) used OMR for the first time and proceeded to use it in the Population Census 1991, Agriculture Census 1983-84, and Economic Census 1986, 2001 & 2003. In 2001 the BBS used OCR technology as it proved to be faster than previous capture methods and they procured scanners and associated experts to prepare forms and test form processing and train BBS staff. Initially the Bangladesh Government planned to use the technology for the population census data capture process but the infrastructure was used further for preparing voter lists & ID cards. The hardware consisted of 4 OCR machines, 2 servers & 31 workstations. Software allowed for the automatic capturing and managing information (data) from forms (questionnaire), which interpreted and verified the forms, then transferred the data to a host system (Server). The presentation further explained the use of OMR for the 2001 and 2003 economic census and outlined good practices in data processing with a focus on batch management. This was followed by successes and specific components needed for the use OMR as well as associated problems. The presentation concluded by emphasizing specific considerations to be taken into account in the use of both OMR and OCR such as environmental conditions, speed, and accuracy.

43. **Bhutan**: The last census was carried out on 30-13 May 2005. The presentation by Bhutan focused on forms processing, data processing and coding hierarchies and discussed the forms used in the 2005 census. The steps involved the checking of forms by enumerators, team supervisors, Gewog supervisor and finally the district statistical officer before being sent to the Office of census commissioner for processing. The presentation also covered the manual data processing methodology used by Bhutan and other factors such as training and number of staff involved. It concluded with a discussion on the different responsibilities for data processing personnel and detailed place and code standards for the region during the population and housing census of 2005.

44. **Brunei Darussalam:** The presentation by the Brunei Darussalam Department of Statistics provided an overview on the main 2 stages of data capture employed by the office currently. The first stage involved house numbering and the second stage was the actual enumeration execution where face to face interviews were conducted by field staff over a 10 day period. The current mode of data capture in the department of statistics is manual key entry. Also discussed were the two main phases of data processing and the statistical packages used. The first phase involves general particulars and one economic activity question and the second phase covered all remaining questions. The splitting of the processing into two phases was not recommended to be used in the next census. Further data processing stages were outlined such as file batch conversion, editing, and coding. The Brunei Darussalam Department of Statistics will consider the practicality in the use ICR Technology for the next census. The office currently seeks expertise in the area of data processing and coding as well as GIS mapping and dissemination.

45. **Cambodia:** The country conducted its census on 3 March 2008. The presentation by the Cambodia National Institute of Statistics discussed the data processing workflow, focusing on the enumeration area batch. It was explained that the batch entails one set of data files for each EA and allows for the processing of data in parallel with data collection. The data processing is split into two stages whereby the first stage involves cleaning and editing the data and the second prepares tables for analysis. Further explanation of the details and hierarchal workflow of the first stage were provided such as data entry, editing, and backup. The second stage included export/import into software packages, tabulation, and adding geographic data such as GPS coordinates. The presentation further explained the role of the personnel involved (administrators, operators, and editors) and concluded with an overview of the hardware and physical space needed for the operation and the data entry directory and coding structure.

46. **China:** The presentation by the National Bureau of Statistics of China (NBSC) discussed their vast experience in the use of OCR technology in two cases of large-volume census data capture (1) the 5th national population census and (2) the second national agricultural census. Long and short form types, data storage, staffing and other organizational aspects were discussed and several of the benefits of using OCR data capture methods were also outlined. During the 5th population census, it was indicated that the number of staff needed to capture the data had decreased from previous years. Sampling quality inspections during and after data capture indicated that OCR capture had effectively reduced errors and improved the quality of data. Also, cost accounting indicated that employing OCR method had reduced overall census costs. During the second national agricultural census, a higher efficiency in data capture was obtained. China will carry out the sixth national population census in 2010, and optical data capture, using OCR will be used again in 2011. Due to the success of the 5th national census and 2nd national agricultural census, and to the fact that the organization and staff are familiar with the technology, its operation, and the management and administration of OCR data capture, OCR will continue to be the main data capture method of choice. The presentation further outlined NBSC strategies that will be taken into account during the sixth national population census in 2010.

47. **India:** The presentation began by outlining general features of the census in India and gave detail on the mode for data capture and processing since 1961 or the past 5 completed censuses. This included details on the census year, population, capture percentages, modes of data capture (OMR, OCR etc.), and the time taken to complete. It was followed by more detailed explanation of the 2001 census as it employed ICR tools and technologies for capture and processing. Important considerations, method selection and consequent action, model conceived for implementation and software workflow details, modules and stages along with specific software customization strategies were outlined for the 2001 census. Further to describing the 2001 process, the results achieved, difficulties experienced and lessons learned were discussed in detail as well as he technology to be used in the 2011 census. ICR will continue to be used and the presentation concluded by explaining some improved efficiencies to the ICR data processing methodology that will serve to aid the 2011 census of India.

48. **Indonesia:** An overview of the data processing system including hardware and software was provided with focus on the mainframe system used in earlier censuses. Further discussion went beyond the use of mainframe to discuss decentralized and distributes data processing and IT department and hardware/software structuring. It detailed the data processing workflow for the 2000 population census and lessons learned and focused attention on the data processing methodology that will be used for the 2010 population census. It concluded with a detailed overview of the infrastructure for the 2010 data processing by providing information on scanners in use, established processing centers and progress reporting mechanisms.

49. **Iran:** The presentation by Iran provided a focus on both the Personal Digital Assistant (PDA) and on Intelligent Character Recognition (ICR). It began with an overview on PDA technology which included information on the use of PDA applications in Surveys and Censuses, general advantages, application specifications, localization and customization, survey specific activity considerations and limitations. This was followed by a presentation on ICR technology which entailed a general overview, reasons for use of the OCR method, accuracy, and different stages of operation. It displayed the process schema used in the office as well as gave an overview on character verification, field verification, and batch and rule verification. The presentation concluded by providing a list of strengths with regard to the use of OCR.

50. **Korea (Republic of):** The NSO of Republic of Korea presented an outline of the Korean census, the environment of census taking in the country, internet survey, e-census system, and data capture and editing covered five major data processing components with regard to their experiences. In 2005 the NSO of Korea executed, for the first time, an internet survey with the objectives of decreasing coverage error, providing a way to get at hard-to-enumerate households, and save costs in data collection. Information was provided on the number and characteristics of respondents, composition and processing of the forms and the advantages and challenges. There is now a plan in place for another internet survey for the upcoming 2010 census. Following was a discussion on the e-Census system which aims to be of low cost and have high efficiency, improve data quality, provide a way to get at hard to enumerate households, decrease coverage error, and shorten data release time. The workflow and function of the e-Census system was outlined

in detail. The presentation concluded with an overview of the data capture and editing methods used by the NSO of Korea with specific detail given to the 2005 ICR process.

51. **Mongolia:** The presentation explained the training provided by other countries to NSO staff on the use of specific software and also discussed the hardware infrastructure in place during the 2000 census. Coding and editing as well as the associated issues observed in their specific 2000 census process were presented. It concluded with a summary of the intended plan under discussion for the 2010 census with regard to data processing. The aim of the NSC is to modern technology for data processing by using OCR/ICR at a small scale for capturing questionnaire information, utilize a local network environment for maximizing workflow, and further extend the use of CSPro for data entry.

52. **Nepal:** The presentation by Nepal provided a focus on their experience with outsourcing. Significant background information is provided on the activities of the Central Bureau of Statistics and on general activities regarding data processing and data management. Due to limited resources, data entry along with coding and editing for the 2001 census was outsourced to two private agencies. Further specific outsourcing issues are detailed with a focus on monitoring and supervision and database processing and management. Within the outsourcing operation and the office, several different software applications are described. The presentation concludes with critical lessons learned and challenges. These challenges are outlined within the different stages of the outsourcing operation.

53. **Philippines:** The National Statistics office of the Philippines discussed the workflow for the population census of 2007. The presentation discussed census form types; number of processing centers; provincial, regional and central processing tasks. It also detailed the data processing workflow of the 2000 census and provided a comparison with the 2007 census. Several main characteristics of data processing differences and similarities were outlined such as data capture strategies, ICR application development, scanners, software and scanning sites (geographic location). The presentation concluded with an overview of the main problems associated with the data processing workflow for both the 2000 and 2007 census in which distinctly different issues were observed.

54. **Sri Lanka:** The presentation by the NSO of Sri Lanka focused on the Data Editing Procedure implemented for the 2001 population and housing census. It detailed their use of the IMPS Software for data entry, data editing, imputation and tabulation. The presentation outlined the 3 main editing methods: range edits, structural and consistency edits. The data processing division and subject matter division steps and workflow were detailed with regard to these methods. The main distinction was that specific attention to be given during the use of IMPS programs for range edits and for structural and consistency edits together. The presentation concluded with a discussion on the graphical output showing an IMPS error print of structural & consistency edits, following with the same type of graphical output for range edits. For the next Census in 2011 the NSO of Sri Lanka will use ICR and has decided to outsource all of the process. The census survey processing system (CSPro 3.3) will be used for data editing. A team will be appointed to develop an automated data coding system for the next Census.

55. **Thailand:** The presentation by the NSO of Thailand focused on the 2010 census preparatory activities. It began with a history of the most current major activities in preparation for the 2010 census. It provided a brief outline of the entire population and housing census execution and activities. This included discussion on issues associated with planning, timing, mapping, questionnaire forms, testing phases executed, data collection methods to be employed, data processing, data dissemination and public relations. It concluded by presenting some of the problems with the 2000 census and in its use of ICR.

56. **Vietnam:** The presentation by Vietnam focused on data processing of the 1999 Vietnam Population Census and gave insight into some of the strategies that will be used in the 2009 census. It gave an overview on the three recent population censuses in 1979, 1989 and 1999. This was followed with a focus on the data processing methods used in the 1999 population and housing census. It outlined software selection, data processing facility infrastructure, data entry, editing and tabulation operations and system management and control. There was also a discussion on data dissemination methods. The presentation concluded by providing some of the strategies that will be applied in the 2009 census.

Session 4: Outsourcing versus in-house Processing

57. In this session, the UNSD made a presentation on outsourcing of specific tasks at different stages of census operations. Most of the National Statistical Agencies responsible for conducting census are not capable of carrying out all the tasks involved in conduct of census. The reasons include (i) lack of necessary technological expertise or equipment at NSO; (ii) the need for improving timeliness and accuracy of the data, (iii) a recognition of the complexity of job; and (iv) the added advantage that the NSO gains access to external expertise and knowledge. A decision on whether or not to outsource should be based on:

    a. Defining the technical needs of the NSO in terms of expected output

    b. Specifying the requirements for the delivery of the output in terms of timeliness, quality assurance, accuracy, confidentiality, etc

    c. An assessment of the market *vis-à-vis* the NSO needs to determine if it would be feasible to undertake the outsourcing

58. Contractor and NSO should have a shared understanding of the requirements of the contract, including objectives, expected outcomes and priorities. Clear Specifications, including standards to be met, are key to ensuring to get what is wanted and that everyone understands what is expected. Specifications should describe in detail the tasks that are the responsibilities of the NSO and for the contractor. Specifications should include detailed milestones with deliverables against which performance should be evaluated. Specifications for the output should also address requirements for timeliness, data confidentiality and security, quality assurance.

59. The discussion focused on the following points:

    a. Costs of outsourcing and duration of the process. It was recognized that outsourcing can improve efficiency but has an impact on the costs.

    b. Accuracy: how to define the relevant indicators and how to measure them? More generally, It was stressed that the NSO should keep the control on the process?

c. Confidentiality: how to include confidentiality requirements in the contract?

Session 6: Introduction to Data Capture

60. This session was devoted to a discussion on the methods of data capture, the relative advantages and disadvantages of the various methods, and issues relating to choice of an appropriate method. The presentation made by the UNSD began by defining “data capture” as a process of converting collected data to a computer interpretable format. It described five main methods of data capture: (i) keyboard data entry, (ii) optical mark recognition/reading (OMR), (iii) optical character recognition/intelligent character recognition (OCR/ICR), (iv) personal digital assistant (PDA), (v) Internet, and revealed the limitations and relative advantages of each method.

61. Choice of method should be part of the overall strategic objective of the census in terms of timeliness, accuracy and cost. The technology used should be decided early in census cycle in order to allow enough time to test and implement the system. When imaging technology is used for data capture, extensive testing is required well in advance of the census.

Session 7: Data Capture: Optical Mark Recognition (OMR)

62. This session consisted of two presentations; one by the UNSD and the other by the representative of the DRS, from UK. The presentation of UNSD mainly dealt with definitions and concepts of the method. OMR is a technology that allows an input device (e.g. imaging scanner) to read hand-drawn marks such as small circles or rectangles on specially designed paper. An OMR works with a specialized document and contains timing tracks along one edge of the form to indicate where the scanner can read for marks which look like black boxes on the top or bottom of a form. The advantages of OMR are that this data capture technology does not require a recognition engine. Therefore: it is fast, using minimum processing power to process forms, and costs are predictable and defined. Conversely, OMR is unable to recognize hand-printed or machine-printed characters. Tick boxes may not be suitable for all types of questions. In that regard, questionnaire design and preparation is critical. Field Operators must take particular care in filling out questionnaires, and training is essential.

63. DRS provided the technical details of the two available OMR technologies, that is, OMR from image and OMR dedicated scanners.

64. During the discussion, a number of issues were raised by the participants such as the printing requirements and the conditions of storage of the paper.

Session 8: Data Capture: OCR / ICR / IR

65. This session consisted of a presentation by UNSD and by DRS and BetaSystem. UNSD presented the definitions and the main difference between the OCR, ICR and IR technologies. OCR gives scanning and imaging systems the ability to turn images of machine printed characters into machine readable characters, while ICR is able to process hand written characters. OCR/ICR has less strict form design compared to OMR. Forms can be scanned through a scanner and then the recognition engine of the OCR/ICR system interprets the images and turn images of handwritten or printed characters into ASCII data (machine-readable characters). Images are scanned and stored and maintained electronically, therefore, there is no need to store the paper forms as long as you safeguard the electronic files. The relative advantages and disadvantages were also highlighted in this presentation. Scanning and recognition facilitate efficient

management and planning for the rest of the processing workload, but this technology is costly and may require significant manual intervention.

66. The presentations by DRS and BetaSystem were focused mainly on form design, hardware/ software requirements, workflow, accuracy, and relative advantages and disadvantages of the three methods.

67. During the discussion that followed the presentations, participants asked precisions about the “false positive” rate, that is to say the characters wrongly recognized by the machine. It was underlined that this parameter is more important that the accuracy rate, because a character not automatically recognized can be proposed for manual recognition, whereas a character wrongly recognized cannot be detected. There was also questions on the possibility of scanning separate sheets, the different storage options, the optimal scanner resolution as well as the costs.

Session 9: Manual Data Entry, PDA, Internet

68. This session consisted of two presentations – one by UNSD followed by presentation by Republic of Korea. UNSD presented the basics of manual data entry. Computer-assisted keyboard data entry usually uses personal computer data entry programs with built-in logical controls. Some of the tasks accomplished by the programs are: (a) verifying that EA codes are valid and copying them automatically from one record to the next; (b) assigning a number to each person in a household automatically and to each household within an EA; (c) switching record types automatically if required; (d) checking that variable values are always within pre-determined ranges; (e) skipping fields if the logic indicates doing so; (f) supporting keyboard verification of the information entered earlier; and (g) generating summary statistics for the operator and the batch. Decision to use manual entry versus automated entry may take into account: timetable requirements, relativities between staff and hardware costs and possibility to implement more sophisticated technology. Where staff costs are low and computing infrastructure is moderate, keyed entry may be the optimal method.

69. UNSD discussed the types of PDAs, as well as key specification features currently available on the market, relative advantages and disadvantages, and the criteria for making choice of the PDAs. It was noted that having extensive training prior to the deployment of PDAs is essential. It is critical that the vendor provide post implementation support – for both technical and hardware aspects.

70. Internet data collection is an alternative for self-enumeration by paper forms. Some countries, such as Canada and Australia, have already implemented such mode for data collection, at the request of the population or within the frame of a government initiative. The main benefits, are accurate and timely data collection and reduction in the processing time. Internet provides rapid availability of clean data for statistical analysis.

Sessions 10: Data Capture: Process Stages

71. UNSD made a presentation to highlight the major issues regarding stages of the census process, from the field operations to the production of an edited microdata file ready for tabulation. The importance of a complete and updated EAs list was underscored. It was also noticed that reporting population aggregates by EA and capturing it quickly gives the potential to produce first results and to control the data capture process.

72. DRS presented its approach to Census Data capture process, including the pre-census planning. The process stages include forms receiving, scanning, recognition, verification, quality assurance/management and logistics issues. All these process stages were presented in detail with illustrative examples from the Sudanese census, the Ethiopian census on scanning process, the Sudanese and Malawi censuses on verification process, and the Tanzanian and Sudanese censuses on logistics. DRS stressed the fact that the largest issue for time and quality is how well the census forms are filled out by the enumerators.

73. Beta Systems, a data processing provider from Germany, presented its approach to the census/surveys data capture process stages, which consists of scanning, recognition (OMR, OCR, ICR), and verifying processes, with emphasis on the census data flow and quality assurance. The presentation gave an example of implementation of this approach for the Nigeria census conducted in 2006.

Session 12: Data Capture: Overview of Major Distributors/ Commercial Suppliers

74. During this session, three presentations were made by the data processing providers DRS and Beta Systems, Kodak, Readsoft and TIS. Each provider gave an overview on its specific solution to census data processing and some concrete examples illustrating its implementation

Session 13: Data Coding

75. UNSD gave a presentation on coding of data for censuses, which covered the basic concepts and definitions. Coding is the process in which census questionnaire entries are assigned numerical and/ or alphanumeric values. The objective is to prepare data in a form suitable for entry into computer and for further analysis by users. Two methods were presented: simple coding, limited to reference to one question on the census form, e.g., birthplace; and structured coding, used for complex topics (e.g. occupation, industry, education, etc.), in which reference may be made to more than one question. Regardless of system used, they all rely on coding indexes. The indexes are lists of typical responses likely to be given on a census form that has associated classification code assigned to them. Coding operations may involve one of the three options: (a) assigning numerical codes to responses recorded in words or in a form requiring modification before data entry; (b) rewriting numeric codes recorded say on a questionnaire to a separate coding sheet to facilitate data entry, and (c) use pre-coded entries on questionnaires which may be used directly for data entry. The appropriateness of closed-ended versus open-ended questions was explored. The relative merits of manual versus computer-assisted versus automatic coding was covered. Finally a few of the common international classification systems (ISIC and ISCO) were discussed.

76. Participants discussed the strategy to adopt regarding coding, especially in countries with many languages as India. The balance between automated and manual coding was discussed and questions were raised about the different kinds of algorithms. Participants requested UNSD to share the knowledge of software packages used in the world.

Session 14: Introduction to Data Editing

77. UNSD gave a presentation on the introduction to data editing, commencing with a definition of the terms used. Editing is the procedure for detecting and correcting errors from data, and Imputation is the procedure of assigning values to missing or

inconsistent data. The presentation described the types of errors typically encountered in the census process – including both content and coverage errors. An illustration was given showing why it is important to edit, especially in terms of overall trend-related distributions of data. Some basic principles of editing were given and the concepts of fatal versus query edits, micro versus macro-editing, and manual versus automated editing were presented. Finally the pitfalls of over-editing were discussed.

78. The discussion highlighted that editing rules must be elaborated by a team composed of demographers, statisticians and IT specialists. Demographers must give and endorse the editing, whereas IT specialists implement these rules into a computer program, and also give rules related to the database structure. It was suggested to organize the edits with a first set of variables in order to disseminate quickly first results. These variables should not be modified by further edits.

Session 15: Concepts and Methods in Data Editing

79. UNSD gave a second presentation on data editing, going into more detail into the notions of within record editing and across record editing. The presentation gave the definition of the different concepts. "Structure edits" check coverage and relationships between different units: persons, households, housing units, enumeration areas. "Validity checks" are performed to see if the values of individual variables are plausible or lie within a reasonable range, whereas "Consistency checks" are performed to ensure that there is coherence between two or more variables. Two different ways of running the edits were presented: the Top-down Editing approach starts by editing top priority variable (not necessarily first variable on questionnaire) and moves sequentially through all items in decreasing priority; the Multiple-Editing approach uses a set of rules that state the relationship between variables. The process of imputation changes one or more responses or missing values in a record or several records to ensure internally coherent records result. Two methods of imputation were presented: Cold Deck and Hot Deck. In the Cold-Deck approach, values are imputed on a proportional basis from a distribution of valid responses (e.g., from previous census). In the Hot-Deck approach one or more variables are used to estimate the likely response based on data about individuals with similar characteristics.

80. The participants discussed the number of variables to take into account in the imputation matrix, and the trade between efficiency and simplicity. For example, the neighbouring households are usually processed together, and are usually close in terms of housing characteristics. This should be kept in mind when designing the edits as the addition of too many variables in the Hot-Deck would entail to take a donor far from the household. Participants also stressed the importance of taking into consideration the national social context in the design of the edits.

Session 16: Data Editing (Practical Exercises)

81. UNSD gave an introductory presentation on the basics of CSPRO, a software package freely available on-line, developed by US Census Bureau, and having three main functionalities: data entry, data editing and data tabulation. Then the presentation showed several examples of implementation of edits using CSPro, and detailed the writing of the codes and running on the sample database. Examples covered basic imputation, control between two variables, control among the members of a household, and example of Hot-Deck imputation.

Session 18 - Data Documentation and Archiving

82. Facilitated by presentations by ESCAP (Ilpo Survo and Imae Mojado), the Workshop discussed the importance of and tools available for documenting and archiving census and household survey microdata sets. The Workshop agreed that the provision of access to microdata had become more important and easier than during the 2000 round of censuses and recommended that statistical offices review relevant legislation and institutional arrangements, and develop explicit policies for the use of microdata. The Workshop emphasized that only appropriately anonymized microdata sets should be made publicly available. Simple removal of the obvious identifiers, such as names and the detailed geographic location, did not reduce the risk of disclosure adequately.

83. The new Microdata Management Toolkit and international standards made the documentation of data sets much easier than previously. The Toolkit allowed combining data and metadata in one user-friendly package, which could be easily distributed and archived. The Workshop agreed that the preservation of microdata for long-term development analysis could only be achieved through proper documentation and preservation of the data sets. The experience ESCAP gained during its recent regional pilot project showed that the tools were developed enough to be adopted as a standard practice for new household surveys and censuses. The documentation could be started during the planning phase of censuses/surveys and was worth of doing for important historical data.

Recommendations & Conclusions

84. Regarding the method of data capture, the workshop recommended that countries carefully assess their capacity and cost factors before opting for any particular technology for their next census. Different methods of data capture were accordingly discussed, including OMR, OCR/ICR, use of PDA and handheld devices, internet and manual data entry. It was agreed that there is no one optimal mode of capture. The use of new technology should not be influenced solely by current trends but rather by national needs.

85. NSOs/census offices should encourage its staff to be proactive in acquiring the relevant knowledge of technologies from contracted consultants or solution providers in order to build capacity.

86. The meeting discussed, at length, of the possibilities of outsourcing as part of the census process in the area of data capture. In this regard participants recommended that decisions pertaining to outsourcing should be taken early enough during the preparatory stages of a census in order to allow time for the biding process, for testing and implementation of technical specifications. The following aspects should be considered before embarking on outsourcing:

    a. An assessment of skills available (including project management) at the national statistical/census office to prepare the tender documents and implement the technology;

    b. Estimated cost with regard to the availability of resources;

    c. National rules pertaining to procurement;

    d. Once the decision to outsource has been taken, the contract should describe precisely the deliverables expected, the time frame, and should include strict confidentiality and security requirements, as well as a quality assurance plan.

87. The workshop recommended that the census data capture process should have a complete quality assurance plan, regardless of the technique used. The different parts of the process should be monitored with few but reliable indicators. In that regard, specific attention should be paid to the proportion of false positive, that is to say the proportion of characters recognized wrongly.

88. The participants discussed different stages pertaining to the processing of census data. They emphasized the importance of early preparation for the planning of data capture activities. In addition, it was recommended that the pilot census should include a test of data capture, data editing, and data tabulation.

89. In developing the whole process of data capture, national statistical/census offices should ensure that edit specifications and codes for open-ended questions are developed in collaboration with subject matter specialists based on pilot tests and previous experience.

90. The participants recommended that Statistical/census offices adopt statistically sound editing and imputation strategies as part of their data processing operations. Caution should be exercised to avoid over-editing. However, it was agreed that there is no editing solution that can identify every error. The participants urged that special attention be given at every stage of the preparation of the census, such as questionnaire design and training of the enumerators.

91. The Workshop recommended that statistical/census offices should increase the value, quality and usability of the conducted population and housing censuses by documenting the respective data sets from the beginning in accordance with international standards and good practices and by ensuring that the documented data sets are archived properly. It noted that the Microdata Management Toolkit provided a suitable framework for the purpose and requested ESCAP and other development agencies to keep countries current in their use in population and housing censuses.

92. The Workshop also recommended that the statistical and census offices should assess their need and requirement for microdata sharing with a view to improving data analysis and policy planning. For this, countries may have relevant national legislation, and their organizational mandates, policies reviewed, and amended where needed, to facilitate safe access to microdata by researchers.

93. The participants expressed, with great appreciation, the benefits gained from the exchange of relevant experiences during the workshop. In addition, countries recommended for enhancement of the knowledge pertaining to census data processing and urged UNSD to collect and disseminate information, for example editing programs, through its 2010 World programme website and knowledge base.

94. NSOs/census offices are encouraged to share experiences, approaches and best practices through technical cooperation and study tours for all aspects of census processes and operations including data capture, coding and editing.

EVALUATION OF THE WORKSHOP

95. Overall, the participants appreciated the workshop’s main focus on new technology for data capture and editing. The participants emphasized that the most useful element of the workshop was sessions regarding Data Processing, and in particular Data Capture sessions. Also they appreciated the sharing of experiences and ideas between participating countries. Prevailing comments suggested more practical and demonstration oriented sessions. Also, participants would prefer publications to CD's,

and they will appreciate better involvement of participants in open discussion. General opinion of the participants is that the workshop was well organized and held in a timely manner considering upcoming census.

ANNEXES

Annex I. Agenda of the Workshop

Annex II. List of participants

Annex I.   Agenda of the Workshop

UNSD-UNESCAP Regional Workshop on Census Data Processing: Contemporary technologies for data capture, methodology and practice of data editing, documentation and archiving

Bangkok, Thailand, 15-19 September, 2008

Agenda

Time	*Topic*	*Responsibility*	*Document*
Monday September 15, 2008			
	Opening		
9:00-9:30	***Registration of participants***		
9:30 – 10:00	**Session 1 – Opening remarks -** welcoming remarks by UNSD, UNESCAP, administrative matters	UNSD (JMD)	
10:00 – 10:15	***Group photo***		
	Review of United Nations Principles and Recommendations for Population and Housing Censuses and the preparation of the 2010 round of censuses in the region Objective: To present revisions of the United Nations Principles and Recommendations for Population and Housing Censuses followed by a round table on the preparation of 2010 round of censuses in the region.		
10:15 – 12:00	**Session 2 – The 2010 World Programme on Population and Housing Censuses /** UN Principles and Recommendations for Population and Housing Censuses, Rev.2- major issues in the revision/ Status of Data Capture in Censuses in the ESCAP region – Presentation by UNSD – Presentation by each participant of the situation in his/her country – General Discussion	UNSD (JMD)	Pres. 1 (UNSD)
12:00 – 13:00	***Lunch***		
13:00 – 14:30	**Session 3 –Country Presentations- Data Processing/Questionnaire** Results of UNSD pre-workshop questionnaire on capture/ editing (country/ regional analysis) Country/ Regional Presentations on Experiences with Data Processing – Presentation of questionnaire results by UNSD – Presentation of the status of data capture in the region by UNESCAP – Presentation by Thailand – Presentation by Nepal – General Discussion	UNSD (CB), UNESCAP, country	Questionnaire Pres. 2 (UNESCAP)
14:30 – 15:00	***Coffee break***		
	Introduction to Data Capture Methods and Outsourcing versus in-house processing Objective: To present an overview of Data Capture management considerations and present and discuss the applications and issues of data capture using Optical Mark Recognition Technology; Optical Character Recognition/Intelligent Character Recognition, Manual Data Entry and provide an overview of different process stages		

Time	*Topic*	*Responsibility*	*Document*
15:00 – 15:45	**Session 4 – Outsourcing versus in-house processing.** Is outsourcing required? How to manage outsourcing. Country examples on outsourcing of data capture – Presentation by UNSD – General Discussion	UNSD (JMD)	Pres. 3 (UNSD)
15:45 – 16:30	**Session 5- Country Presentations on Data Processing (I)** – Presentation by Afghanistan – Presentation by Vietnam – General Discussion	Presentations by Countries	Pres. (Country) Pres. (Country) Pres. (Country)
Tuesday September 16, 2008			
8:30 – 10:00	**Session 6 – Introduction to Data Capture** Methods of data capturing, advantages and disadvantages of each method, issues for consideration when choosing the method. – Presentation by UNSD – Presentation by India – General Discussion	UNSD (CB)	Pres. 4 (UNSD)
10:00 – 10:30	***Coffee break***		
10:30 – 12:00	**Session 7 - Data Capture: Optical Mark Recognition** Construction/Design Characteristics, Hardware and Software Requirements and Scanning/Storage, Advantages and Disadvantages; overview of the major commercial suppliers. – Presentation by UNSD – Presentation by Expert (DRS) – General Discussion	UNSD (CB), Presentation by Experts	Pres. 5 (UNSD) Pres. B (Expert)
12:00 – 13:00	***Lunch***		
13:00 – 14:30	**Session 8 - Data Capture: Optical Character Recognition/Intelligent Character Recognition/ Intelligent Recognition** Construction/Design Characteristics, Hardware and Software Requirements and Scanning/Storage, Advantages and Disadvantages; overview of the major commercial suppliers. – Presentation by UNSD – Presentation by Experts (DRS – BetaSystem) – General Discussion	UNSD (CB), Presentation by Experts, Country Presentations	Pres. 6 (UNSD) Pres. C (Experts)
14:30 – 15:00	***Coffee break***		
15:00 – 16:30	**Session 9– Data Capture: Manual Data Entry/ PDA-Handheld-computers/Internet** Different technologies/processes in data collection using handheld devices (e.g. PDAs) and Internet. – Presentation by UNSD – Presentation by Republic of Korea – General Discussion	UNSD (CB), Country Presentation	Pres. 7 (UNSD) Pres. (Country)
Wednesday September 17, 2008			
8:30 – 10:00	**Session 10 – Data Capture: Process Stages** Scanning, Recognizing, and Verifying Processes associated with data capture and Quality assurance/ management system for data capture and logistic issues as well as how to balance timeliness versus quality – Presentation by UNSD – Presentation by Expert (DRS – BetaSystem) – General Discussion	UNSD (CB), Presentation by Experts	Pres. 8 (UNSD) Pres. D (Experts)
10:00 – 10:30	***Coffee break***		

Time	***Topic***	***Responsibility***	***Document***
10:30 – 12:00	**Session 11- Country Presentations on Data Processing (II)** – Presentation by Iran – Presentation by Indonesia – Presentation by Bangladesh – Presentation by Philippines – General Discussion	Country Presentations	Pres. (Country) Pres. (Country) Pres. (Country) Pres. (Country)
12:00 – 13:00	***Lunch***		
13:00 – 14:30	**Session 12 – Data Capture: Overview of Major Distributors/Commercial Suppliers** – Presentations by Commercial Providers – General Discussion	UNSD (CB)	Commercial Providers Presentations
14:30 – 15:00	***Coffee break***		
	Data Coding Coding is art of preparing data in a form suitable for entry into computer to facilitate analysis. In this connection, the objective of the session to present an overview of different methods of coding.		
15:00 16:30	**Session 13 – Data Coding** Coding systems: Manual (clerical), Computer-assisted, Automatic coding (automatic coding is usually partial) or combination of more than one system. Coding systems in light of the data collection and capture methods planned for the census. Coding of occupations, industry and educational characteristics: At what level of classification? Adapting the international classifications for national use and importance of maintaining international comparability and nationally over time. Coding indexes – Presentation by UNSD – General Discussion	UNSD (JMD)	Pres. 9 (UNSD)
Thursday September 18, 2008			
	Data Editing Objective: Editing is the procedure for detecting and eliminate errors from data. The objective of the session is to present an overview of the concepts and methods and discuss the application and issues.		
8:30 – 10:00	**Session 14 – Introduction to Data Editing** Types of Errors (Coverage + Non-response + Content- questionnaire, enumerator, respondent, coding, data entry, etc,). What is editing (concepts of check, control, correct)? Why Edit (give examples of edited and unedited output tables to illustrate potential biasing)? Pitfalls of over-editing. General description of methods of how to edit and how to impute, concepts of manual and automatic edits – Presentation by UNSD – General Discussion	UNSD (JMD)	Pres. 10 (UNSD)
10:00 – 10:30	***Coffee break***		
10:30 – 12:00	**Session 15 – Concepts and Methods in Data Editing** Within Record Editing – Concepts of validity and consistency checks, examples of both population and housing edits, example of how edit specifications done, concept and methods of imputation – Presentation by UNSD – General Discussion	UNSD (JMD)	Pres. 11 (UNSD)
12:00 – 13:00	***Lunch***		
13:00 – 14:30	**Session 16 - Data Editing** – Exercises on CSPRO – General Discussion	UNSD (JMD)	Pres. 12 (UNSD)
14:30 – 15:00	***Coffee break***		
15:00 – 16:30	**Session 17- Country Presentations on Data Processing (III)**	Country	Pres. (Country)

Time	*Topic*	*Responsibility*	*Document*
	– Presentation by Mongolia – Presentation by Bhutan – Presentation by Sri Lanka – Presentation by Malaysia	Presentations	Pres. (Country) Pres. (Country) Pres. (Country)
Friday September 19, 2008			
	Country Presentations on Data Processing (Cont.) **Final Report, Recommendations & Conclusions**		
8:30 – 10:00	**Session 17- Country Presentations on Data Processing (IV)** – Presentation by Cambodia – Presentation by Brunei Darussalam – Presentation by China	Country Presentations	
10:00 – 10:30	***Coffee break***		
	Data Documentation and Archiving		
10:30 – 12:00	**Session 18 - Data Documentation and Archiving (I)** Different methods and practices for documentation and archiving census data. – Presentation by UNESCAP – General Discussion	UNESCAP	Pres. 13 (UNESCAP)
12:00 – 13:00	***Lunch***		
13:00 – 14:30	**Session 18 - Data Documentation and Archiving (II)** Different methods and practices for documentation and archiving census data. – Presentation by UNESCAP – Presentation by Country – General Discussion	UNESCAP	Pres. 14 (UNESCAP)
14:30 – 15:00	***Coffee break***		
	Final Report, Recommendations & Conclusions		
15:00 – 16:00	**Session 19 - Final Report, Recommendations & Conclusions** – Final Report, Recommendations & Conclusions: review and adopt report, conclusions and recommendations (Final report lead by Rapporteur, evaluation of Workshop)	UNSD	Final Report

Annex II: List of participants

UNITED NATIONS - DEPARTMENT OF ECONOMIC AND SOCIAL AFFAIRS

ECONOMIC AND SOCIAL COMMISSION FOR ASIA AND THE PACIFIC

UNSD/ESCAP Regional Workshop on Census Data Processing: Contemporary technologies for data capture, methodology and practice of data editing, documentation and archiving

15-19 September 2008

Bangkok

AFGHANISTAN

Mr Mohammad Sami Nabi, President of Field Operations and Sampling Department, Central Statistics Organization, Kabul

BANGLADESH

Mr Jatan Kumar Saha, Systems Analyst, Bangladesh Bureau of Statistics, Ministry of Planning, Dhaka

BHUTAN

Mr Khandu Dorji, Senior Statistical Officer, National Statistics Bureau, Thimphu

BRUNEI DARUSSALAM

Ms Marilyn Linggi Teo Lai, Assistant Director of Statistics, Department of Statistics, Department of Economic Planning and Development (JPKE), Prime Minister's Office, Bandar

Ms Halimatussaadah Perkasa, Statistician, Department of Statistics, Department of Economic Planning and Development (JPKE), Prime Minister's Office, Bandar Seri Begawan

CAMBODIA

Mr Pen Socheat, Demographic Statistics, Bureau Chief of Demographic Statistics, Phnom

CHINA

Mr Li Xiru, Deputy Director-General, Department of Population and Employment Statistics, National Bureau of Statistics, Beijing

Mr Xia Yuchun, Director of Division of General Affairs, Computer Center, National Bureau of Statistics, Beijing

INDIA

Mr Devender Kumar Sikri, Secretary, and Census Commissioner, Office of the Registrar General & Census Commissioner, India, New Delhi

Mr Ramesh Chander Sethi, Additional Registrar General, Office of the Registrar General, India, New Delhi

Mr Mohan Singh Thapa, Joint Director, Office of the Registrar General & Census Commissioner, India, New Delhi

Mr Chinmoy Chakravorty, Joint Director, Office of the Registrar General, India, New Delhi

INDONESIA

Mr Ichwan Ridwan Tandjung, Head of Division, Data Communication Network, Badan Pusat Statistik, Jakarta

ISLAMIC REPUBLIC OF IRAN

Ms Somaye Ahangar Saryazdi, System Analyst, Statistical Center of Iran, Tehran

MALAYSIA

Mr Mazlan Sulong, Principal Assistant Director, Department of Statistics, Putrajaya

Ms Zaitun Mohd Taha, Statistician, Department of Statistics, Putrajaya

MONGOLIA

Mr Munkhbadar Jugder, Senior Officer In-Charge of Population and Housing Census Issues, National Statistical Committee of Mongolia, Ulaanbaatar

MYANMAR

Ms Soe Soe Aung, Director, Department of Population, Ministry of Immigration and Population, Nay Pyi Taw

Mr Sein Myo Aung, Deputy Director, Department of Population, Ministry of Immigration and

Population, Nay Pyi Taw

Mr Myo Thwin, Assistant Director, Department of Population, Ministry of Immigration and Population, Nay Pyi Taw

Mr Win Myint, Staff Officer, Department of Population, Ministry of Immigration and Population, Nay Pyi Taw

NEPAL

Mr Bharat Raj Sharma, Statistical Officer, Central Bureau of Statistics, Kathmandu

PHILIPPINES

Mr Valentino Abuan, Director, Information Resources Department, National Statistics Office, Manila

REPUBLIC OF KOREA

Ms Sung Ok You, Deputy Director, Population Census Division, Population & Social Statistics Bureau, Korea National Statistical Office, Daejeon

Ms Juwon Lee, Deputy Director, Information System Development Division, Statistical Information Service Bureau, Korea National Statistics Office, Daejeon

Ms Soorin Hwang, Statistician, Population Census Division, Population & Social Statistics Bureau, Korea National Statistical Office, Daejeon

SRI LANKA

Mr Hettiarachchige Rohana Dias, Deputy Director, Department of Census and Statistics,

Mr Weerasiri Hewage Priya Wasantha Weerasiri, Senior Systems Analyst/Programmer, Department of Census and Statistics, Colombo

THAILAND

Ms Chitrlada Touchchai, Computer Technician, National Statistical Office, Bangkok

VIET NAM

Mr Van Cam Mai, Deputy Director, Department of Population and Labour Statistics, General Statistics Office, Hanoi

UNITED NATIONS SECRETARIAT

Mr Jean-Michel Durr, Chief, Demographic Statistics Section, Statistics Division, DESA, New York

Mr Charles Brigham Reese, Census Cartographer, Demographic Statistics Section, United Nations Statistics Division, DESA, New York

UNITED NATIONS BODY

United Nations Population Fund (UNFPA)	Ms Wassana Im-Em, Assistant Representative, UNFPA, Bangkok

OTHER ENTITIES

Beta Systems Software AG	Mr Richard J. Lang, Director Consulting International, Augsburg
DRS Data Services Limited	Mr Andy Tye, International Manager, Milton Keynes
Kodak (Singapore) Pte. Limited	Mr Susheel John, Business Manager - South East Asia Document Imaging, Graphic Communications Group, Singapore
Kodak (Thailand) Limited	Mr Paiboon Fuangkawinsombut, Account Manager, Document Imaging, Graphic Communications Group, Bangkok
Read Soft Asia Sdn Bhd.	Mr Joakim Dahl, Solution Specialist, Kuala Lumpur
Top Image Systems (Asia Pacific) Pte.Ltd	Mr Eli Shoshani, Managing Director, Shanghai Mr Ido Schechter, CEO, Top Image Systems, Tel Aviv

OBSERVERS

Ms Kanjana Phumalee, Statistics Technical Officer, National Statistical Office, Bangkok,

Ms Vilairat Anantaphrut, Statistics Technical Officer, National Statistical Office, Bangkok, Thailand

Ms Angkana Hatthakitkosol, Computer Technical Officer, National Statistical Office, Bangkok, Thailand

Ms Oranuch Hutajata, Socio-Economic Statistician, National Statistical Office, Bangkok,

Ms Pannee Pattanapradit, Socio-Economic Statistician, National Statistical Office, Bangkok, Thailand

ESCAP SECRETARIAT

Ms Noeleen Heyzer	Executive Secretary
Mr Shigeru Mochida	Deputy Executive Secretary
Mr Herve Berger	Chief of Staff
Mr Richard Kalina	Secretary of the Commission
Mr Srinivas Tata	Special Assistant to the Executive Secretary

Mr Pietro Gennari	Director, Statistics Division
Ms San Yuenwah	Statistical Analysis and Publication Coordinator
Ms Haishan Fu	Chief, Statistics Development Section, Statistics
Mr Ilpo Survo	Chief, Statistical Information Services Section, Statistics Division
Mr Jan Smit	Statistician, Statistics Division
Mr Joel Jere	Statistician, Statistics Development Section, Statistics Division
Mr Eric Hermouet	Statistical Information Systems Officer, Statistical Information Services Section, Statistics Division
Mr Andres Montes	Statistician, Statistics Development Section, Statistics Division
Ms Imae Ann Mojado	Statistician, Statistics Division

Mr Wei Liu	Associate Statistician, Statistics Development Section, Statistics Division
Mr Christian Stoff	Associate Statistician, Statistical Information Services Section, Statistics Division
Ms Zeynep Orhun	Associate Statistician, Statistical Information Services Section, Statistics Division