UNSD-UNESCAP 공동 주최

센서스 데이터 자료처리 및 내용검토관련 워크샵

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

2008. 10

통계청 인구조사과

< 순 서 >

Ⅰ 회의 개요

1. 회의명

센서스 자료입력 및 내용검토에 관한 워크숍

2. 회의 기간

2008. 9. 15 ~ 19 (출장기간 : 2008. 9. 14 ~ 9. 21)

3. 개최지

태국 방콕 UNESCAP

4. 회의 참가국

16개국 43명 참가

(중국, 태국, 필리핀, 말레이시아, 몽골, 이란, 인도, 인도네시아, 부르나이, 아프가니스탄, 네팔, 스리랑카, 베트남, 부탄, 방글라데시, 캄보디아, UNFPA, 기타 자료입력업체 등)

5. 참가자 및 활동

- 인구조사과 류성옥 사무관, 전산개발과 이주원 사무관, 인구조사과 황수린 주무관
- 2010 인구주택총조사 자료처리를 위한 각국의 경험 공유 및 토론 참가
- 한국의 2005 센서스 및 자료처리 방식에 대한 발표

Ⅱ 회의 일정

1일차	9.15(월)	등록 및 오프닝 session 1 :2010 UN 센서스 권고안 개정부분 자료처리분야 토의 session 2 :각 국의 2010 센서스 준비상황 및 자료처리과정 발표 I session 3 :데이터 입력에 대한 외주(Outsourcing)방안	The United Nations Building Rajadamnern Nok Avenue Bangkok 10200 Thailand
2일차	9.16(화)	session 4 :데이터 입력방식에 대한 개괄 OCR, ICR 자료 기술, 각 국의 데이터 입력에 관한 경험 공유 및 토의 session 5 : OCR, ICR 에 관한 토의	
3일차	9.17(수)	session 6 : PDA,인터넷 자료처리에 관한 토의 session 7 : 데이터 코딩(Data coding)	
4일차	9.18(목)	session 8 : 센서스 자료 내검(Editing)에 대한 토의 데이터 내검 실습 데이터 자료처리과정에 대한 각국의 발표	
5일차	9.19(금)	데이터 자료처리과정에 대한 각국의 발표 session 9 : 최종 발표, 결론	

Ⅲ 주요 발표 내용

1. Session 1 : 2010 World Population and Housing Census Programme

❑ 2000년 센서스

27개 국가는 센서스를 수행하지 못하였는데 그 이유는 정치적인 문제나, 재정문제, 적절치 못한 계획이나 관리, 숙련된 직원들이 부족해서였다.

❑ 2010년 전 세계 센서스 프로그램은 다음과 같은 세 가지 목표를 가지고 있다.

1. 센서스를 수행하는데 있어 전국가적인 원칙과 추천에 대한 동의
2. 2005~2014년 사이에 센서스를 수행하도록 촉진
3. 센서스 결과를 적정한 시기에 공표하도록 지원함

❑ 첫 번째 목표

센서스의 원칙과 추천에 대한 두 번째 버전이 2007년 3월에 채택되었다.

- 자료처리 프로세스
- 국제적인 자문 프로세서의 확장
- 지역과 국제적인 자문을 통한 점진적인 과정
- 미팅, 이메일 교환, 온라인 회의 등

○ 원칙과 추천- 주요한 변화점

- 자료에 기초한 의사결정을 위한 결과물 생산 강조
- 결과물 생산을 위한 통합된 통계시스템 개발

	비핵심에서 핵심으로	핵심에서 비핵심으로
가구원	- 마지막에 태어난 아이의 출생일자 - 지난 12개월간의 출생자 - 지난 12개월간의 가구원 중 사망자 - 지난 12개월간 태어난 아이 중 사망자 - 출생국 - 도착년도, 기간 - 장애 정도	- 근무시간
주택	- 음식을 조리하는데 사용되는 연료 - 식수의 주요공급원 - 가구에서 사용되는 정보통신기기들	- 건축년도, 기간 - 사용되는 방수 - 임대료

○ 다른 핵심 주제들
- 예전 거주장소
- 과거 특정기간 거주 장소

○ 분리된 핵심주제들
- 화장실의 종류
- 쓰레기 처리실태

○ 거처의 종류에 대한 새로운 분류법
- 전통적인 주택의 개념과 현대적 주택의 개념의 분명한 차이
- 집단 거처의 분류범위 확대

❑ 대안적인 접근

- 대안적인 접근 노력의 도입

* 전통적인 센서스에서 등록 센서스나 롤링센서스로의 전환

- 외부와의 계약 체결
- 센서스 funding

- 품질보증 (품질통제가 아닌)
- GIS, Geo- 코딩을 이용한 발전된 센서스 맵핑
- 조사하기 어려운 홈리스나 원주민들에 대한 집중

❑ 두 번째 목표

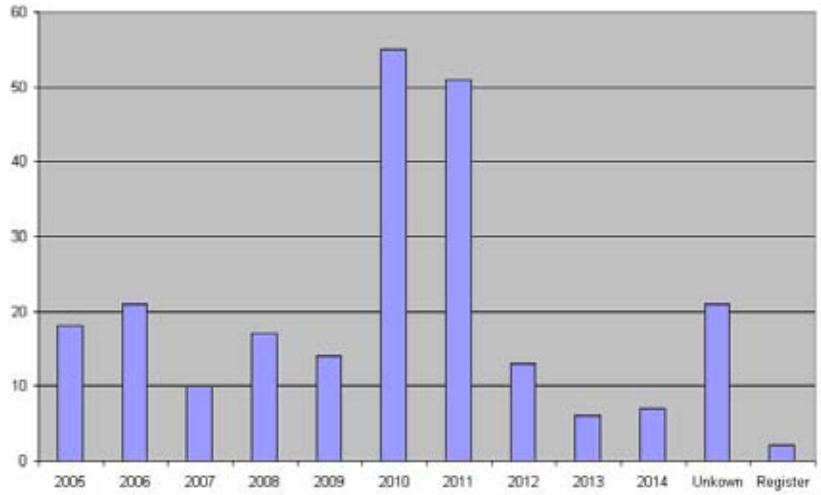
○ 2010 센서스를 수행하기 쉽도록 국가들을 독려함

- 66개국은 2008년 말까지 센서스를 수행하기로 약속하였다.
- 21개국은 아직 센서스 계획을 수립하지 못했다.

○ 도움이 필요한 주요한 문제는 다음과 같다.

- 자금
- 준비와 계획
- 지도화
- 관리와 품질보증
- 자료입력 특히 스캐닝
- 데이터 프로세스
- 자료배포와 분석
- 평가와 사후조사
- 정치적인 문제들을 포함한 기타 문제들

○ 센서스 수행 계획연도



❑ 세 번째 목표

○ 각 국가들이 센서스 결과를 공표할 수 있도록 지원함
부적절한 데이터 공표와 분석은 지난 센서스에서 취약점이었음

- 센서스 데이터 공표와 분석에 대한 핸드북 제공
- 센서스 데이터 공표와 분석에 대한 워크숍 개최
- 센서스 데이터 공표에 대한 소프트웨어 개발 : *CensusInfo*

○ 2010 센서스를 위해 UNSD의 지원 활동

* 파트너십
  - UNSD는 UNFPA와 파트너십을 강화하고 UN의 지역회의, 다른 국제적, 지역적 기구와 협력한다.
  - 국가의 요구를 모니터하고 수집하고, 트레이닝하고 기술적 원조를 제공하며 정보를 교환하기 위해

* 모니터링
  - 센서스를 수행하는 것과 관련있는 국가의 활동을 모니터링함
  - 센서스 관리자들과 접촉함
  - 지속적으로 업데이트하고 UNFPA와 지역적 회의들과 협조함
  - 2010 센서스 프로그램을 웹사이트와 뉴스레터 등을 통해서 정보를 공유함

* 기술적인 가이드라인을 제시함
- Principles and Recommendations
- 핸드북, 센서스 주제별 가이드라인
- GIS를 이용한 센서스 지도화
- 국제적 이민 등 측정
- 센서스에서 경제적 특성파악
- 자료내검
- 평가, 사후조사
- 새로운 방법에 대한 자료제공

* 교육
  - 센서스를 수행하는데 있어 문제되는 기술적 이슈들에 대해 다음과 같은 워크숍을 수행함
  - 2006: Principles & Recommendations: WS in Mozambiq ue, Mali
  - 2007: 센서스 지도화와 관리 : 4 WS in Zambia, Thailand, Trinidad, Morocco
  - 2008: data processing:
  - 자료입력 (internet, scanning, PAD, inc. outsourcing)
  - 내검
  - 품질보증, 위험관리
  - 2009 이후: 공표, 분석, 평가 (PES)
  - 온라인 교육 개발

* 경험 공유와 제안
  - 워크숍은 국가들의 발표와 결실 있는 경험을 공유하고자 한다.
  - 워크숍은 실습을 포함한다.
  - 현실화 되고 있는 기술과 방법을 강조한다(예를 들면 *GIS 등).*
  - 교육목적으로 방문하는 것을 지원한다.

2. Session2 : 워크숍 전 사전조사 결과

❑ 질문서의 목적

○ 각 나라 수준에서 자료처리활동을 더 잘 이해하기 위하여

○ 자료처리의 기법과 기술의 효과적인 사용에 대한 더 많은 협력제공을 위한 토론회와 각 나라의 경험들을 공유하기 위하여

○ 워크숍 관리와 더 많은 활동들을 지원하기 위하여

○ 특이한 자료처리 방법의 사용에 필요한 정보와 기술 훈련을 이해하기 위하여

❑ 자료입력: 총조사 자료 입력 방법

○ 총조사 자료입력에 사용되는 방법

- 수동으로 입력
- OMR
- OCR/ICR

○ 몇몇의 나라는 PDA's와 인터넷을 이용에 관심이 있음

❑ 자료 입력: 각 나라에 사용된 스캐너와 특징들

○ Kodak i610 80 - 320 images/sheets per min.(필리핀)

○ Kodak 3510 200 dpi (60 - 75 ppm) 300 dpi: (40-50 ppm) (싱가포르)

○ Fujitsu M4099D ~90 ppm[simple] to 180 images per minute [duplex], up to 400 dpi(말레이시아)

❑ 자료 입력: 외주 입력

○ 수동자료 입력에 대한 관심과 더불어 자료입력이 늘 외주 입력되는 것은 아니다.

○ Oracle이나 CSPro 같은 데이터베이스 관리 시스템은 자체에 자료 입력이나 내검, 코딩 방법들을 포함하고 있다.

○ OMR과 OCR/ICR에 대한 관심과 함께 자료 입력과정은 부분적으로 혹은 전체적으로 외주 입력 방식을 취하고 있다. (예: 방글라데시, 인도네시아, 스리랑카)

○ 휴대용 장비들/인터넷의 사용이 제시되고 있다.(예: 싱가포르, 이란)

❑ 자료 입력: 문서화 방법과 정책들

○ 많은 나라는 조사표 저장을 위하여 전자적 수단을 사용한다. 몇몇 나라는 전자적 형태와 종이 형태, 두 가지로 조사표를 저장한다.

○ 몇몇 나라는 일정기간 동안 조사표 보관을 요구하는 법이 있다.

○ 인쇄된 종이 형태의 저장은 문제를 일으킨다. 그것들은 공간을 차지하고 일정한 기간이 지나면 손상될 수 있다.

❑ 자료 내검: 직업 분류를 위한 코딩

○ 각 나라는 직업, 산업, 교육을 코드화 한다.

- 직업: 대부분의 나라가 ISCO를 사용하지만 몇몇 나라는 자국의 특별한 체계와 함께 사용함
- 산업: 대부분의 나라가 ISIC를 사용하지만 몇몇 나라는 자국의 특별한 체계와 함께 사용함
- 교육: 대부분 ISCED를 사용하지만 몇몇 나라는 자국의 특별한 체계와 함께 사용함
- 인종: 또한 코딩이 사용되는 중요한 분류로 언급된다.

❑ 자료 내검: 수동 또는 자동 코딩

○ 대부분의 경우 코딩은 수동으로 행해진다. 그러나 몇몇 나라는 수동과 자동 두 가지의 방법을 다 사용하기도 한다.

- 자동 코딩 소프트웨어는 자체 개발하거나(이집트), Oracle처럼 상업적으로 생산된 것을 사용하거나(레바논), 사적 계약자에 의해 개발되거나 통계청 직원에 의해 구현(모나코)되기도 한다.

○ 대부분 나라는 총조사 처리 단계의 한 부분으로써 내검 시스템을 가지고 있다.

- 조사표 안에 표현된 오류감지 시스템

- 타당성 체크와 일관성 체크
- 레코드나 도표 안에서의 교차 체크

○ 많은 경우, 무응답 대체를 위하여 CSPro, IMPS, SPSS, Oracle 등의 소프트웨어와 함께 수동의 방법들이 사용된다.

○ 대부분의 나라들은 자료 입력 프로그램과 함께 SPSS STATA, 일괄 내검 프로그램 등의 통계적 소프트웨어를 이용하여 자동화된 루틴을 만들어 낸다.

○ 몇 몇의 나라는 내검 시스템을 자체 개발해 사용하기도 한다.

(예: 방글라데시, 한국, 필리핀)

❑ 조사원 채용 및 훈련

○ 훈련에 걸리는 기간(전국)

	전체	수동	OMR	PDA
자료 입력	5일~1달	1주	12일	28일
자료 코딩	5일~2주	3일	5일	7일
내검	5일~3주	3일	5일	7일
무응답 대체	1일~3주	1일	-	-

❑ 각 나라별 품질관리 과정: 자료처리 단계와 관련이 있는

○ 캄보디아

- 자료처리 조정자, 품질관리팀, 검증자, 총관리자, 검증 양식, 생산 양식, 품질관리 게시판, 그룹별 소규모 품질관리 모임, 특별한 개인들을 목표로 하는 훈련 등

○ 말레이지아

- 자료입력단계: 자료 해석과 검증에 대한 표본 체크
- 자료코딩단계: 표본화된 양식으로 체크
- 무응답 대체: 일관성 체크와 개요 표 생산

- 제표: 견본 표 생산

○ 필리핀

- ICR을 기본으로 한 자료 입력
- 자료 입력 및 수동 코딩
- 한정적 빈도수(Marginal Frequencies).

3. Session3 : 아웃소싱 (Outsourcing)

❑ 센서스 운영에서 아웃소싱 : 왜 아웃소싱 하는지

○ 기술적인 전문가나 통계청의 장비가 부족하므로
○ 시간과 데이터 질의 정확성을 제고하기 위해
○ 이로 인해 통계청은 그들의 핵심 업무에 집중할 수 있음
○ 아웃소싱된 업무는 복합하기 때문
○ 통계청은 아웃소싱을 통해 외부전문지식과 전문가를 얻을 수 있음

❑ 아웃소싱할 것인가와 무엇을 아웃소싱 하는지

아웃소싱할지 여부의 결정은

○ 기대되는 결과의 관점에서 기술적 필요사항을 정의하고
○ 신속성, 데이터 질의 보장, 정확성, 보안성의 관점에서 세부 요구사항을 점검함
○ 시장의 평가를 측정해 보아야 함

❑ 아웃소싱을 위한 기술적 요구사항 정의

○ 아웃소싱에서 요구되는 결과와 산출물에 대해서 명확한 정의를 내려야 함
○ 프로젝트의 목적, 달성하고자 하는 결과와 산출물, 소요시간 등을 포함

❑ 세부 요구사항

- ○ 계약자와 통계청은 계약의 요구사항과 목표 기대 효과 및 우선순위를 이해하고 공유해야함
- ○ 만족할 만한 표준을 포함하여 모든 사람이 이해하고 동의할 만한 명확한 세부사항 설계가 필요함
- ○ 통계청과 계약자의 의무를 포함하여 세부사항을 합의해야함
- ○ 결과에 대한 세수사항은 신속성 자료보완, 자료의 질 보증 등을 작성해야함

❑ 시장에서의 평가

- ○ 통계청은 아웃소싱계약이 체결되기 이전에 시장에 대한 명확한 정의가 필요함
- ○ 아웃소싱된 작업에 대한 기술적 평가
- ○ 프로젝트에 대한 잠재적 경쟁자에 대한 평가
- ○ 아웃소싱된 비용의 추정
- ○ 통계청이 아웃소싱을 감당할 수 있는지에 대한 평가
- ○ 입찰에 대한 준비에 대해 도움

❑ 아웃소싱 조달 과정

- ○ 입찰과정은 매우 긴 시간이 걸림
- ○ 투명한 경쟁과정을 거친 계약자의 선택
- ○ 조달입찰 과정은 국가마다 다르지만 일반적으로 다음을 포함한다.
  - 제한시간이 있는 공고규칙
  - 무엇이 아웃소싱 되는지에 대한 자료
  - 통계청은 입찰경험을 공유하기 위해 다른 정부부서와 협조를 얻을 수 있음

❑ 계약자의 선택

- ○ 증명된 자료에 의해 계약자를 선택하는 것이 바람직
- ○ 세부요구사항을 만족해야하고 단지 비용뿐아니라 신뢰성 , 정확

성, 시의성 등 다른 사항들도 고려해야함

○ 계약자들이 평가받기 이전에 기준을 공고해야 함

○ 예상 계약자들은 경쟁 테스트를 통과해야 한다. 예를 들면 다른 회사들이 조사표를 스캔하여 그들의 실행능력을 시험받아야 한다.

❑ 계약의 종류와 아웃소싱

○ 장단점을 가진 다른 종류의 계약이 있다.

- 완전한 계약- 계약자가 모든 작업을 아웃소싱해서 전담함
- 혼합계약- 여러 회사와 계약해서 각자 맡은 분야가 있음
- 하위계약- 계약자가 하위 업체와 재계약하여 업무를 수행함

○ 혼합계약의 장점은 각 분야의 계약된 작업을 잘할 수 있는 여러 기업들을 고용할 수 있다는 점이나, 관리와 책임소재 측면에서 문제가 있다.

❑ 계약의 행정사항

○ 통계청과 계약자의 의무를 세부적인 법적효력이 있는 문서 계약을 통해 협의하는 것이 필요하다.

○ 계약은 관리와 전반적인 운영에 있어서 통계청의 명확한 권리

○ 통계청의 다른 부서나 다른 계약자와 계약위반에 대한 페널티

○ 계약수정을 가능하게 할 유연성

❑ 결론

○ 상황의 평가에 의한 아웃소싱 결정

○ 계약의 명확한 세부화 검토

○ 테스트 시스템의 중요성

○ 관리와 아웃소싱을 모니터하는데 있어서 통계청의 역할

○ 다른 경험으로부터 배우게 되는 점

4. Session4 : 자료입력 개관 (Data capture)

❑ 자료입력이란 무엇인가?

자료입력이란 센서스를 통해 수집된 정보를 컴퓨터로 인식 가능한 형태로 변환시키는 작업을 의미함

(UN Principles and Recommendation rev.2)

❑ 자료입력방식

- ○ 키보드 데이터엔트리
- ○ OMR
- ○ OCR / ICR
- ○ PDA
- ○ 인터넷
- ○ 위의 여러 가지 방식의 조합

❑ 키보드 데이터 엔트리

- ○ 센서스 조사표로부터 얻은 응답코드를 컴퓨터에 수작업으로 입력하는 것
- ○ 시간과 비용이 고려되어야하며 보다 정교한 기술의 수행이 필요함
- ○ 문자화된 응답을 분류카테고리에 넣는 작업이 필요함

* 장점

단순한 소프트웨어와 성능이 낮은 하드웨어도 가능함

인건비에 따라 다르지만 비용이 적게 든다.

입력 PC를 센서스 이후에 다른 용도로 사용할 수도 있다.

* 단점

많은 인력이 필요함

자동화된 데이터 입력방식에 비해서 시간이 많이 필요함

데이터 입력과정에서 잠재적인 오류 발생률이 높음

작업이 개인능력에 따라 달라지므로 작업의 표준화가 어려움

❑ OMR

- ㅇ OMR은 조사표를 스캐닝 하는 방식으로 키보드가 아닌 컴퓨터로 조사표가 읽혀진다.
- ㅇ OMR은 특별하게 디자인된 종이에 tick-box에 표시한 응답내용을 읽게 된다.
- ㅇ 스캔된 응답지는 코드로 변환된다.
- ㅇ 손으로 쓰인 응답지는 손으로 입력되거나 computer-assisted 방식을 통해서 코드화 된다.

* 장점

데이터의 정확성 제고
키보드로 입력하는 것보다 빠름
장비비용이 저렴함
상대적으로 설치하고 운영하기 간편함
많은 국가에서 이용되어왔던 방식임

* 단점

조사표 디자인에 제한이 있음
조사표 종이와 잉크에 제한이 있음
조사표인쇄와 절단과정에서 정확성이 요구됨
응답마크에 정확히 표시 돼야 하고 필기도구도 제한됨
텍스트는 인식 불가함

❑ OCR / ICR

- ㅇ OCR/ICR은 조사표 전체와 문자를 인식할 수 있는 기술이다.
- ㅇ OCR 기술은 프린트된 문자만 인식하기 때문에 필기체는 수작업으로 입력되거나 코드화 되어야 한다.
- ㅇ ICR은 프린트된 문자 뿐 아니라 수기로 작성된 문자도 인식한다.

* 장점

OMR방식보다 조사표 제한이 덜하다.

자동화된 조사표 인식 과정으로 시간이 절약된다.

조사표가 데이터 파일화되어 보관하기 용이하고 다음에 사용할 때 조사표 재생이 가능함

어떤 수기 응답들은 자동적으로 코드화되어 데이터의 질을 제고함

* 단점

장비비용이 비쌈 (정교한 하드웨어와 소프트웨어 필요)

시스템을 위한 IT 요원이 필요함

인식에러를 피하기 위해 수기로 작성한 문자는 정자로 쓰여야 한다.

문자를 대체하는 과정이서 데이터 질이 저하됨

인식률제고의 문제는 인식률의 질과 비용의 상반관계에 있다.

❑ PDA

○ 센서스 조사표의 내용이 PDA에 저장되어 스크린에 나타남

○ 데이터가 바로 저장되며 통계청의 데이터베이스로 전송된다.

* 장점

자료입력측면에서 수기입력 비용을 줄일 수 있다.

자료유효성 검증측면에서 다시 검증하는 비용을 줄임

실시간으로 논리적 유효성을 검증하기 때문에 시간이 절약되고 논리적 오류가 줄어든다.

빠른 센서스 결과가 가능함

* 단점

프로세스 과정을 세팅하는데 시간이 많이 걸리며 세밀한 테스트가 필요함

조사원이 PDA를 다룰 줄 아는 사람인지 테스트 하기위해 행정적 절차가 필요함

조사원에게 PDA사용 교육이 필요함

조사기간에 배터리가 닳지 않도록 여유분이 필요함
장비가 고장 날 위험성이 있음

❑ 인터넷 조사

○ 인터넷을 통해 센서스 자료를 수집하는 예가 증가하고 있음 (그러나 이 방법은 다른 조사방법과 항상 보완적으로 사용되고 있다.)
○ PDA와 같이 종이조사표를 PC에 다운로드 하는 방식은 아님
○ 조사표를 작성하기 위해서는 비밀번호가 필요함
○ 내부전문가의 부족으로 인터넷조사방식을 위해서는 아웃소싱 되는 경우가 많음

* 장점

조사표를 제작하거나 자료입력 하는데 비용이 절약됨
조사원이 조사하기 어려운 지역이나 집단에 대해 조사하기 좋은 방식임
관계없는 질문에 대해서 자동 필터링됨
상호 유효성 체크가 되므로 데이터 질이 제고됨
데이터 입력과 에디팅 과정이 간편해지므로 빠른 데이터 결과도출이 가능함

* 단점

인터넷 접속을 위해서는 컴퓨터가 필요함
응답자가 응답시 문제가 생길 때 바로 처리가 불가능함
높은 수준의 보완시스템이 필요함
모든 응답자가 인터넷을 사용하는 것이 아니므로 평행적인 처리과정이 필요함
누락과 두 번 제출하는 것에 대한 체크가 필요하다.
시스템을 세팅하고 적절하게 운용하기 위해서 비용과 시간이 많이 든다.

❑ 자료처리방식을 선택함에 있어서 제기되는 문제들

- ㅇ 방식은 국가의 환경에 따라 달라짐
- ㅇ 처리방식의 선택은 신속성, 정확성 경제성 원칙에 의해 결정되어야 함
- ㅇ 자료처리시스템의 선택과 기술은 센서스 사이클에서 초기에 결정되어야 함
- ㅇ 테스트 하기에 충분한 시간과 시스템이 갖추어져야함
- ㅇ 내부에서 전문가가 부족할 경우 아웃소싱할 가능성 있음
- ㅇ 자료입력이 PDA나 인터넷조사로 이루어질 경우 보다 강화된 테스트가 요구됨
- ㅇ 디자인이나 조사표 종이의 질은 자료입력방식과 직결됨

❑ 조사표 설계에 대한 조언

- ○ 조사표안에 담을 조사항목 수를 고려하여야 함
- ○ ticks를 담는 통이 있는 곳 가까이에 사전 인쇄된 코드를 둔다.
- ○ 자료 입력 과정의 속도를 고려할 것 - 마크나 ticks를 가능한 한 많이 사용하는 것이 바람직하다.
- ○ 제거할 색깔을 적절하게 정의하고; 빠른 인식을 위하여 표시 마크 (registration marks)를 사용하여라.
- ○ 수집된 정보를 담을 수 있는 일정한 양식을 유지하여라.
- ○ 제목, 라벨, 명령 등이 붙은 ticks와 마크의 시야를 가리지 말 것
- ○ 질문에 대한 답을 다른 페이지에 두지 말 것
- ○ 개방형 질문 사용을 자제하여라.

❑ 스캐닝으로부터 좋은 결과를 얻는 방법

○ 적절한 종이의 질 선택

- 조사표 종이는 1 square당 80g보다 무거워야 스캐너로 인한 손상을 막을 수 있다.

○ 믿을만한 인쇄기(printing press) 선택

○ 탈락색깔(drop out color)을 고려한 적절한 잉크 사용

❑ 자료 수집 및 처리 시 고려할 사항

○ 현장 작업

- 현장 작업자는 선택된 자료입력과정에 대하여 기본적인 지식을 갖고 있어야 한다.

- 현장 작업 시 고려할 사항들

  · OCR의 오독 이유들

  .. 먼지, 접힘, 구겨짐으로 인한 나쁜 상태의 조사표

  .. 점, 장식적인 필법, 구부림 등으로 인한 불필요한 선이 가미된 문자들

  .. 조사표의 완성도와 일관성에 대한 점검

○ 자료 처리요원 훈련

- 요원들에게 필요한 교육 실시는 입력된 자료의 질과 유효성을 보장한다.

- 교육 내용

  · 하드웨어와 소프트웨어 등 장비 설치에 관한 내용

  · 기초적인 소프트웨어 지식

  · 스캐너 작동 절차들

  · 자주 일어나는 문제(고장)에 대한 간단한 수리 및 해결 방법

○ 조정 단계

- 조정 단계는 자료의 이미지가 부분적이거나 자료가 없는 경우 생성된 파일의 질을 보장하기 위하여 실시된다.

  · 값 검증 단계

  · 공백 조정

  · 부족한 조사내용

5. Session 5. : OCR, ICR

❑ 개념

○ OCR

- 출력된 문자를 인식해서 데이터화하는 스캐닝, 이미지화 시스템
- 프린트된 문자 이미지는 스캔된 이미지에서 비트맵으로부터 이미지를 추출한다.

○ ICR

- 수기로 작성된 문자를 인식해서 데이터화하는 스캐닝 이미지 시스템
- 수기로 작성된 문자 이미지도 스캔된 이미지에서 비트맵으로부터 이미지를 추출한다.

❑ OCR과 ICR의 차이

○ OCR은 OMR 보다 덜 정확하지만 ICR보다는 정확하다

○ ICR은 높은 데이터 질을 위해서는 반드시 내검이 필요하다.

○ 조사표 형식

- OCR / ICR은 OMR보다 조사표 디자인에 구애를 덜 받는다.
- 타임트랙이 필요 없으며 등록 마크를 가진다.
- ICR은 알파벳이나 숫자를 적을 때 각각의 박스 내에 적어야 한다.

○ OCR

- OCR / ICR은 타이밍 트랙이 필요 없으며 조사표 디자인 구성에 덜 민감하다.
- 드랍 컬러의 사용은 스캐너 출력물의 사이즈를 줄여주고 정확성을 제고시킨다.
- ICR 기술은 등록마크를 네 코너에 사용하여 이미지를 인식한다.

❑ OCR / ICR 스캐너와 소프트웨어

○ 조사표는 스캐너를 통해서 스캔되며 인식엔진으로 이미지는 데이터화되고 수기와 프린터 화된 문자가 ASCII데이터로 변환된다.
○ 사용자는 OCR작업 없이도 스캔될 수 있다.
○ 분당 85-160시트 속도로 인식된다.

❑ OCR / ICR 데이터 저장

○ 이미지는 스캔되고 저장되어 전자적으로 관리된다.
○ 종이조사표를 저장할 필요가 없다.
○ OCR / ICR기술로 이미지는 스캔되고 광학 장치로 읽힌다.

❑ 이상적인 OCR / ICR 정확성

○ 정확성

- 데이터 입력원에 의해 입력되는 경우 완벽한 튜닝이후 OCR / ICR 정확성과 비슷한 99.5%수준이다.

○ 에디팅 이후에 99.9%까지 정확성이 높아짐

❑ OCR / ICR의 장점

* 장점

이미지를 이용한 인식은 높은 수준의 데이터 셋을 생산할 수 있다. OCR / ICR인식기계는 출력된 문자나 손으로 쓴 문자를 인식할 수 있다.

스캐닝과 인식기능은 효율적인 관리와 데이터 처리과정 업무 부담을 계획할 수 있게 한다.
데이팅 후 빠른 재생이 가능하다.

* 단점
인식기술이 비싸다.
주요한 수기 작업이 필요할지 모른다.
자료입력에 부가적인 작업이 필요할 수 있다.
ICR은 사람이 손으로 쓴 문자에 제한이 많다.
수기로 쓰거나 프린트된 문자는 각각 박스에 쓰여야 한다.
흘려 쓴 글씨체가 있는 경우에 효과적이지 않다.

❑ OCR / ICR 제기되는 문제점

- ○ OMR의 문제점과 비슷하다
- ○ 인식엔진에 의한 자료처리과정의 시간고려
- ○ 개발 비용

❑ IR의 개념

- ○ 수기체나 필기체, 흘려 쓴 글씨체를 이미지를 데이터로 인식할 수 있는 시스템
- ○ 비트맵으로부터 이미지를 뽑아내서 인식함
- ○ 흘려 쓴 글씨체를 인식할 수 있다는 것은 이 기술의 독보적인 점이다.
- ○ 인식과정에서 사용되어 결과의 정확성을 제고한다.

- ○ 이미지가 모호한 경우 문자의 분할요소가 문자를 인식하는데 도움을 준다.
- ○ 모든 필기체의 문자를 8개의 요소로 분석한다.

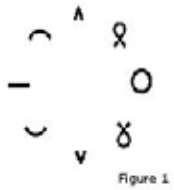


Figure 1

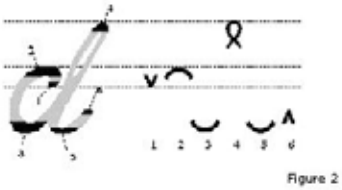
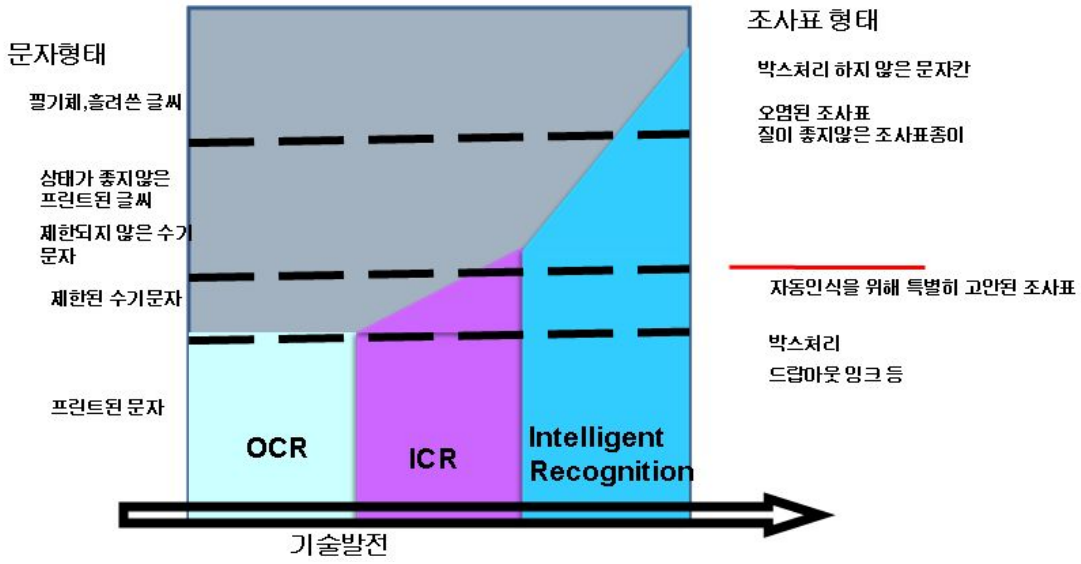


Figure 2

❑ 기술 발전수준



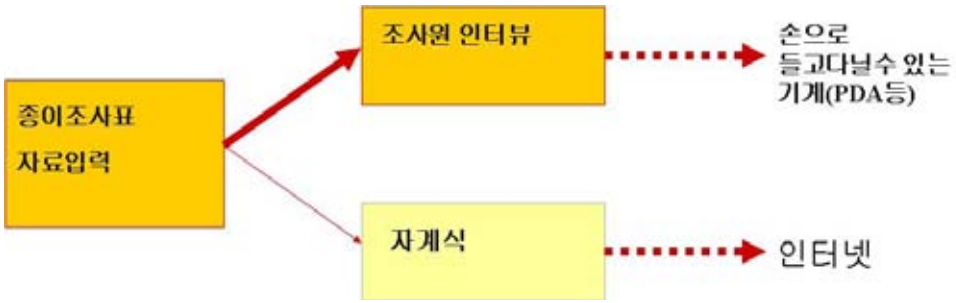
❑ 주요한 기술 공급자들

- Top Image Systems (TIS) (http://www.topimagesystems.com)
- ReadSoft (http://www.readsoft.com)
- Teleform (http://www.intelliscan.com/TeleForm1.htm)
- 스캐너 공급자

  Fujitsu, Canon, Bell & Howell, Kodak

6. Session 6 : PDA/인터넷 자료입력

❑ 개요



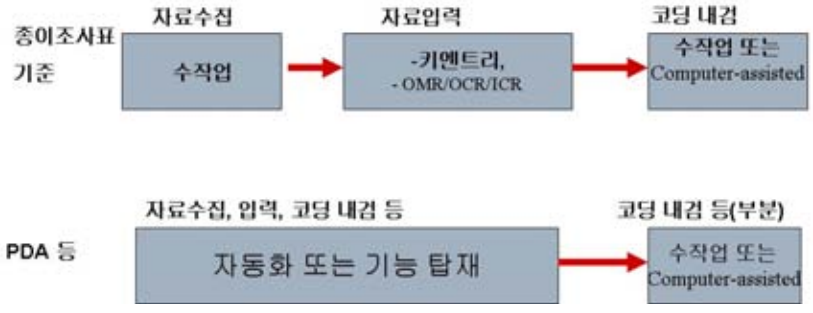
❑ 종이조사표의 문제점

- ○ 느리고 시간이 걸린다.
- ○ 다시 값을 입력하는 것은 비효율적이고 오류가 날 확률이 많다.
- ○ 하나의 자료처리 과정에서 여러 가지 종이표를 제출해야한다.
- ○ GIS와 같은 데이터를 수집하기 어렵다
- ○ 데이터 통합성과 신빙성에서 문제가 있다.

❑ PDA와 소형 컴퓨터

- ○ PDA(Personal digital assistant)는 손에 들고 다닐 수 있는 컴퓨터 이며 소형 혹은 palmtop 컴퓨터라고 불리는 것이다.
- ○ 손에 들고 다닐 수 있는 컴퓨터는 PDA보다 더 많은 기능을 가지고 있으며 서로 호환이 가능하다.

❑ 종이조사표 와 PDA 조사의 업무흐름도



❑ PDA : 중요한 기능

- ○ 동시에 데이터 입력과 코딩이 가능하다
- ○ 두 가지 방식으로 데이터 입력이 가능한데 객관식처럼 고르는 방법과 문자를 타이핑하는 방식이 있다.
- ○ 현재 질문에 답하지 않으면 다음질문으로 넘어갈 수 없도록 설계할 수 있다.
- ○ 실시간으로 집계가 가능하다.
- ○ 각 조사원들이 조사하는데 필요한 시간을 측정할 수 있다.

❑ PDA의 주요 기능들

- ○ Pocket 컴퓨터와 같이 작동한다.
- ○ RAM, ROM등을 사용할 수 있으며 기계에 다운로드 받을 수 있다.
- ○ 데이터는 현장에서 디지털 포맷으로 수집된다.
- ○ 데이터를 서버로 바로 업로드 할 수 있다.

❑ 연계 옵션들

- ○ GPRS/CDMA
- ○ 데이터 스토리지를 증가시킬 수 있다
- ○ 부가기능들
  - 카메라, GPS, 스마트 카드리더 등

❑ 다른 중요한 기준

- ○ 실시간 데이터 업로드 vs 배치 데이터 업로드
- ○ 데이터 접속비용
- ○ 쉽게 데이터입력이 가능한지, 한손에 쉽게 들고 다닐 수 있는지
- ○ 배터리 수명, 배터리 충전기계 등

❑ 현장에서 기계를 관리하는데 발생하는 문제들

○ 현장에서 기계를 어떻게 관리할 것인가
○ 데이터 보안문제

❑ 장비 보안문제

○ 기계를 업데이트하는 문제
○ 기계가 고장 났을 때 문제 등

❑ 부가적 기능들

○ 기계는 GPS기능도 가능하다
방문하고자 하는 곳의 지리적 좌표를 알려주고
데이터 입력의 자취를 알 수 있으며
조사원의 위치를 알 수 있으므로 신변보호기능도 가능하다.
○ 핸드폰 기능과 카메라 등 기능도 가능하다

❑ PDA 비용

○ 포켓 PC : 300~1200달러
○ 비용은 사용자인터페이스와 업그레이드 비용 등에 달려있으며
규모의 경제에 따라 달라짐(많이 구입하면 비용이 싸짐)

❑ 인터넷 조사 방법

○ 왜 인터넷 조사를 하는가
- 국민들이 원하기 때문(프랑스)
- e-government 정책(캐나다)

○ 장점
정확성과 시의성 있는 데이터 입력
자료처리과정에서 시간이 절약됨
빠른 데이터 크리닝과 분석
분석틀을 제고시킴

다른 부서에서 재사용이 가능함
효율성을 증진시킴-계획자 등 관련 인원을 감소하고 비용을 절감시킴

❑ 캐나다의 경험

○ Government On line 동기
○ 2006 센서스
- 캐나다 우체국을 통해 73% 조사표가 발송되었다
- 인터넷조사 개발과 시스템 처리 과정 등은 사기업으로 아웃소싱 되었다.
○ 목표 : 인터넷으로 20%응답

❑ 2011년을 위한 유럽 센서스 기구의 시험조사

시험조사는 두 가지 중요한 측면을 포함한다.
○ 인터넷조사와 PDA와 같은 손으로 들고 다닐 수 있는 컴퓨터
○ 그들은 센서스 수행자가 PDA를 들고 다니면서 국가의 주소 등을 업데이트 하기를 원했고 종이조사표를 제출하라고 했지만 실패한 가구를 추적해서 인터넷조사를 하는 것이다.

❑ 결론

○ PDA를 이용한 센서스는 양적, 질적으로 많은 국가에서 사용되고 검증되었다
(호주, 캐나다, 브라질, 말레이시아, 뉴질랜드 등)
○ 다양한 옵션 등을 추가해서 사용할 수 있으며 실제 조사 이전에 조사원에게 기계에 대해 충분히 훈련을 시켜주는 것이 중요하다.

7. Session 7 : 데이터 코딩

❑ 코딩이란 무엇인가?

○ 총조사 조사표에 기재된 사항을 숫자나 알파벳으로 바꾸는 과정이다.

○ 코딩 목적은 컴퓨터 입력이나 사용자 분석을 위하여 자료를 적당한 형태로 바꾸는 것이다.

○ 총조사 조사표의 각각의 질문에 가능한 응답을 만들고 이 응답들을 숫자나 알파벳에 대등시키는 것이다.

❑ 코딩 방법론 (Coding methodologies)

○ 단순한(Simple) 방법

- 직선적 방법
- 하나의 질문에 조회하는 제한적인 방법(예: 출생지)

○ 구조적(Structured) 방법

- 복잡한 주제(문항)에 사용됨(예: 직업, 산업, 교육정도 등)
- 조회를 위하여 하나 이상의 질문들이 만들어짐
- 운영자를 위하여 코딩 규칙들은 구조화된 코딩시스템으로 만들어질 수 있음

○ 한정적(Bounded) 방법

- 코드가 지정되기 전에 다른 수준의 상세함을 얻고자 할 때 사용됨
- 보통 주소부분을 위하여 사용됨
- 코더는 넓은 지역(16개 시·도, 광역시 등)에서 탐색을 시작하여, 분류코드를 얻을 수 있는 좁은 지역(시군구, 읍면동, 거리 등)으로 옮겨 감

❑ 코딩 색인표(Coding indexes)

○ 시스템이 사용됨에도 불구하고, 그들은 코딩 색인표에 의존함

○ 색인표는 총조사 질문항에 주어진 전형적인 응답들의 명단임.

응답자들이 전형적으로 대답한 것에 기초를 둔 전형적인 응답들의 명단은 중요하다. 그것은 단순히 분류구조 내에서 범주만 담고 있지 않으며, 응답자들이 분류용어로 대답하는 것이 아니라 일상의 언어로 대답한다는 사실을 반영한다.

○ 더욱이 그들은 응답들을 여러 가지 분류 구조 내에 나타낼 수 있다.

○ 이 색인들의 질은 훌륭함으로 그들을 만들기 위한 시간과 노력은 과소평가되지 않아야 한다.

○ 색인들은 고정적이지 않으며 때로는 새로운 응답들에 부합되도록 새롭게 보완되어야 한다.

❑ 코딩작업의 종류

코딩 작업들은 세 가지 선택중 하나와 관련이 있다.

○ 자료 입력 전에 단어로 기록되거나, 변형을 요구하는 응답들을 숫자 코드로 할당한다.(예: 지리적 위치, 직업, 산업 등)

○ 자료 입력을 손쉽게 하기 위하여 조사표에 기록된 응답들을 숫자코드로 다시 쓴다.

○ 바로 자료 입력을 위하여 조사표에 사전에 코드화된 기입사항들을 사용

❑ 코드의 종류

○ 사전 코드화된 응답들

- 폐쇄형 응답에 더욱 좋음; 폐쇄형 질문에 있어서 질문항목이나 응답항목에 부여된 번호는 입력 시 코드로 사용한다.

- 사전-코드화된 응답들을 늘이기 위하여 총조사 조사표에는 숫자나 알파벳 코드를 사용하여야 한다.
- 코딩 범주는 상호 배타적이어야 한다.
- 장점: 코드를 개발하기 쉬우며, 시간을 절약할 수 있다.
- 단점: 개방형 질문에는 사용할 수 없다.

○ 사무실 코딩

- 모든 총조사 조사항목이 사전-코드화 될 수 없으며, 개방형 질문에 대해서는 사무실 코딩이 필요하다.
- 응답의 전체 범위가 알려지지 않아, 조사시점에 코드화 할 수 없을 때 조사가 끝난 후에 코딩한다.

○ 개방형 질문

* 장점

- 총조사 기획자가 제시한 단어들로 이루어진 응답들 대신에 응답자들이 선택한 단어들로 자신들을 표현할 수 있다.
- 특히 직업과 같은 복합적 개념을 갖는 항목에 적합하다.
- 연구자들은 응답자들이 그 질문에 대하여 실제로 어떻게 생각하는지를 바로 볼 수 있다.
- 여러 다른 관심을 가진 여러 분석자들이 같은 질문에 대한 여러 다른 응답들로부터 중요한 정보를 얻을 수 있다.

* 단점

- 여러 다른 응답자들이 서로 다른 관점에서 같은 질문에 응답함으로 그들의 응답들은 비교할 수 없다.
- 총조사에서 개방형 질문들은 측정오차의 근원이 된다.
- 총조사 코더(기획자)는 분석을 시작하기 전에 개방형 응답들을 부문별로 코더화 해야 한다. 코딩은 비슷한 대답을 한 응답자들을 함께 묶는 것과 관련이 있다.

○ 코딩과 관련된 문제들

- 총조사에 있어서 모든 조사항목이 사전-코드화 될 수는 없다.
- 응답자가 제공하는 기본 정보와 코딩 명단에 있는 코드와 연결하여 적절한 코드를 부여할 훈련받은 요원이 필요하다.
- 개방형 질문은 응답의 범주를 잘 모르기 때문에 늘 기타항목을 포함하고 있다.
- 자주 사전에 정해진 코드로 연결될 수 없는 많은 문항들이 있을 수 있으며, 그러므로 그러한 응답들은 나중에 사무실에서 코드화 된다.

❑ 코딩 시스템

○ 문자그대로의 자료가 컴퓨터 내검과 제표에 실용적이지 않기 때문에 코딩이 필요하다.

○ 단어나 문장으로 응답한 항목은 아래에 열거된 방법으로 코드화 된다.

- 수작업
- 컴퓨터의 도움을 받은 코딩
- 자동 코딩
- 위에 열거한 방법들의 조합

○ 수작업

- 코딩 요원들이 코드 색인표나 부호집을 보고 응답들을 연결시킴
- 그들은 자료 입력 후에 수동으로 코드를 부여한다.
- 장점: 간단함
- 단점
  - 지루하며, 코딩요원들은 불분명한 항목도 코드를 부여하여 과다한 코딩(over-coding)이나, 편이(bias)가 생길 수 있음
  - 다른 종류의 코딩보다 많은 오류가 발생함

○ 컴퓨터의 도움을 받은 코딩

- 코딩 요원을 돕기 위하여 컴퓨터가 사용됨
- 연합 코드가 데이터베이스 안에 저장되며, 코딩 작업동안 조회할 수 있다.
- 입력 요원이 컴퓨터 터미널에 앉아 코딩요원의 도움 없이 코딩 시트를 보며 입력할 수 있다.
- 실질적인 수행:
  - 코딩 요원은 응답항목의 각 단어 중 몇 개의 문자를 입력한다.
  - 컴퓨터는 적절한 코딩 색인표로부터 연결되는 명단(matching list)을 보여 준다.
  - 코딩 요원은 그 명단으로부터 부합한 색인 목록(entry)을 선택한다.
  - 컴퓨터는 부합한 색인 목록(entry)에 대응하는 코드를 자동적으로 기록한다.
- 예를 들면 “가금 사육자”는 “가 사”처럼 약어로 입력할 수 있다.
- 장점
  - 비교적 효과적이다.
  - 더 질이 좋은 자료를 생산하기 위하여, 자료 처리 요원을 지도하기 위하여 많은 코딩 규칙들을 시스템 안으로 편입시킬 수 있다.
  - 특히 구조화된 코딩에 적당하다.
- 단점
  - 비교적 복잡하며, 개발하기 위하여 시간과 비용이 든다.

○ 자동 코딩

- 전산화된 알고리즘이 입력된 문장으로 된 응답들을 인간의 도움이 없이 대부분의 경우에 맞는 코드를 부여한다.
- 전형적으로 채점 기법(scoring mechanism)과 관련이 있다. 하나

의 응답항목이 코드와 연결되기 전에 특별한 점수가 요구된다.

- 일치율(matching rates)은 변수의 종류와 사용된 알고리즘에 달려 있다.
- 점수가 어떤 수준 이상일 때에 그 응답은 받아들여지며 자동 코딩이 이행되었다 한다.
- 점수가 어떤 수준 이하일 때는 대개 사람의 관여가 필요하다.
- 장점
  - 빠르고, 고효율, 고품질이 가능하며 특히 구조화된 코딩에 적합함
- 단점
  - 복잡하고 비용이 많이 들며, 연관 알고리즘(matching algorithms)이나 색인표에 결점이 있는 경우 계통 오차(**systematic errors**)가 발생할 수 있다.

❑ 코딩 기술(mechanics)

○ 통계청은 자주 총조사와 관련 있는 조사, 둘 다에 사용되는 몇몇 조사항목(예: 출생지, 언어, 민족/인종, 시민권)에 대하여 공통의 코드 명단을 만들어 사용한다.

○ 예를 들면, 공통의 코딩표(coding scheme)에서 “장소”는 지리적으로 다른 수준을 표현하기 위하여 계층적인 3자리 숫자 코드로 이루어진다. 즉 첫째자리 숫자는 지리적으로 가장 넓은 수준을, 셋째자리 숫자는 지리적으로 가장 좁은 수준을 표현하고 있다.

○ 공통의 문제점은 일이나 인종과 같은 변수들의 정의가 다르거나 변하여 총조사와 다른 조사에서 서로 다르게 사용되는 것이다. 그러므로 통계청은 이에 대한 정책 개발이 필요하다.

○ “단순코딩”을 위하여 통계청은 각 질문에 가능한 응답 코드 명단을 만들어야 한다. 예를 들면, 남자는 1, 여자는 2로 혹은 비경제활동인구 중 주부-0, 학생-1, 은퇴자-2, 유아-3, 연로-4, 연금수령자-5, 기타-7로 코드 명단을 만들 수 있다.

- “구조화된 코딩”을 위하여 각 나라는 자기 나라에 맞게 고치거나, 바로 적용할 수 있는 많은 국제 분류 시스템들이 있다.

(a) 국제 표준 산업 분류, ISIC Rev. 4

코드 종류	수준	범주	코드
두 자리 숫자 코드	중분류	식료품 제조업	10
세 자리 숫자 코드	소분류	곡물가공품, 전분 및 전분제품 제조업	106
네 자리 숫자 코드	세분류	곡물가공품 제조업	1061

(b) 국제 표준 직업 분류, ISCO-88

코드 종류	수준	범주	코드
두 자리 숫자 코드	중분류	방문·노점 및 통신 판매 관련직	53
세 자리 숫자 코드	소분류	방문·노점 및 통신 판매 관련 종사자	530
네 자리 숫자 코드	세분류	노점 및 이동 판매원	5305

❑ 코딩 오류의 원천

○ 코딩 규칙들은 불충분할 수 있다.

○ 코딩 규칙들이 부적절하게 적용될 수 있다.

○ 코딩은 부수적인 일이기 때문에 양질의 코드 작업하기 어렵다.

○ 총조사에서 코딩 작업은 대규모이기 때문에 관리하기 어렵다.

8. Session 8 : 내검 (Data editing)

❑ 목표

○ 내검은 데이터로부터 결함을 찾고 오류를 수정하는 작업이다.

○ 무응답대체는 값이 없거나 일관성이 없는 자료에 값을 주는 과정이다.

○ 이 session의 목표는 내검과 무응답의 개념, 정의와 제기되는 문제들에 대해 논의 해보는 것이다.

❑ 센서스 과정에서 일어날 수 있는 오류들

○ 커버리지 에러 (Coverage error)

- 지도나 조사구의 불완전성, 부정확성
- 모든 가구를 포함하는 것에 대한 실패
- 중복조사 누락
- 방문객이나 잠깐 머무는 사람들에 대한 잘못된 계수
- 조사이후에 센서스 자료를 잃어버리는 것

○ 내용에러 (Contents error)

- 조사표 디자인 에러
- 조사원의 에러
- 응답자의 오류
- 코딩오류
- 데이터 입력오류
- 에디팅오류
- 자료화오류

❑ 센서스 과정에서 두 가지 종류의 오류

○ 한 가지는 다음 단계로 진행하는 것을 막는 오류이고
○ 다른 한 가지는 유효하지 않거나 일관적이지 않은 결과지만 논리적으로 다음 단계에 문제는 없는 오류이다.
○ 첫 번째 오류는 모두 수정되어야 하지만 두 번째 오류는 가능한 한 많이 수정되면 좋다.

❑ 내검의 목적 : 왜 내검을 하는지?

○ 내검은 자료를 깔끔하게 해서 분석하기 용이하게 하기 위함이다.
○ 오류의 종류와 근원을 찾아내기 위함
○ 현재 센서스와 다음 센서스 자료의 질을 제고하기 위함
○ 오류를 찾아내는 것 뿐만 아니라 오류의 원인을 찾아내는 것 또한 중요한데 이는 적절한 수정방법과 전반적인 데이터 질을 향상시키기 위해서이다.

❑ 어떻게 내검 할 것인가?

<표1 : 연령, 성별에 따른 2010 인구 -내검 되지 않은 자료와 내검된 자료>

미상(not reported)을 어떻게 처리할 것인가?

	Unedited data				*Edited data*		
Age group	*Total*	*Male*	*Female*	*Sex Not reported*	*Total*	*Male*	*Female*
Total	4,147	2,033	2,091	23	4,147	2,043	2,104
Less than 15 years	1,639	799	825	15	1,646	809	837
15 to 29 years	1,256	612	643	1	1,260	614	646
30 to 44 years	727	356	369	2	729	358	371
45 to 59 years	360	194	166	0	362	195	167
60 to 74 years	116	54	59	3	116	55	61
75 years and over	34	12	22	0	34	12	22
Age Not reported	**15**	**6**	**7**	**2**			

○ 연령 미상과 성별 미상을 응답자에서 알 수 있는 비율대로 똑같이 분배하는 방법이 있다.

예를 들면 23명의 성별 미상자에 대해 알려진 대로 (2033/4147)*23=12가 남자가 되며 나머지 11은 여자가 된다.
비슷하게 15세 연령미상도 알려진 연령그룹비율로 분배한다.
이런 방식은 만약 미상인구가 많다면 편향을 띨 수 있다.
왜냐하면 응답자와 무응답자의 특성은 매우 다르기 때문이다.

○ 이보다 더 발전된 방식은 다른 변수를 고려해서 분배하는 것인데 예를 들면 배우자 관계, 자녀 수 등을 고려해서 연령, 성별 미상을 분배하는 방식이다.

<표2>

Age group	***Numbers***		***Percent***	
	2010	2000	2010	2000
Total	4,147	3,319	100	100
Less than 15 years	1,639	1,348	39.5	40.6
15 to 29 years	1,256	902	30.3	27.2
30 to 44 years	727	538	17.5	16.2
45 to 59 years	360	200	8.7	6
60 to 74 years	116	89	2.8	2.7
75 years and over	34	25	0.8	0.8
Age Not reported	**15**	**217**	**0.4**	**6.5**

○ 미상이 있는 경우 다른 문제는 추세의 분석에 영향을 미친다는 것이다.
○ <표2>에서 만약 미상을 고려하지 않는다면 15-29세의 비율은 27.2%에서 30.3%로 증가하게 된다. 그러나 미상을 분배하고 나면 추세가 달라지는데 28.7%에서 29.3%로 소폭 증가하게 된다.

❑ 내검의 원칙

일반적으로 내검시스템은

- ○ 명백한 오류가 있는 경우만 수정되어야 하고 내검은 최소화 해야 한다. (Fellegi - Holt Principle)
- ○ 가능하면 오류를 찾는 것과 수정하는 것 둘 다 자동화 되어야 한다.
- ○ 다른 통계청의 내검과 일관성이 있어야 한다.
- ○ UN이나 다른 국가들과 내검 기준이 공조되어야 한다.

❑ 치명적 오류(Fatal error)와 의심되는 오류(Query error)

- ○ 내검의 종류

  치명적 오류(Fatal error) : 분명히 오류가 있으며 원인을 규명해야 하는 오류

  의심되는 오류(Query error) : 오류가 있다고 추정되는 오류

- ○ 치명적 오류는 유효하지 않거나 빠진 데이터가 들어있는 것이며 비일관적인 오류도 여기에 포함된다.

- ○ 의심되는 오류는 주관적인 데이터 범위밖에 들어있는 자료에 대한 것이며 같은 조사표의 다른 값에 비해서 상대적으로 값이 높거나 낮은 경우에 해당한다.

- ○ 치명적인 오류는 해결되어야 하나 의심되는 오류는 옳은 값을 찾기 매우 어렵고 오류를 찾아내는 작업을 하는 효과가 치명적 오류에 비해 매우 적다.

- ○ 의심되는 오류에 대해서는 시험조사에서 충분히 연구해 본 후에 비용을 고려해서 내검을 할 것인지 결정해야 한다.

❑ 마이크로 내검과 매크로 내검

- ○ 마이크로 내검

  각각 데이터의 유효성과 일관성을 체크하는 것이며 가구내의 자료에서 연관성을 체크하는 것이다.

○ 매크로 내검

통합된 데이터가 합리적인지 체크하는 것이다

예를 들면 특정연령이 보고되지 않았다면 그 연령에 대해서 무응답대체를 해야 한다.

그러나 매크로 내검은 무응답대체는 전반적인 연령분포 쏠림현상을 야기하지 않는다.

❑ 자료입력방법이 내검에 미치는 영향

○ 키엔트리, PDA,인터넷 방식은 현지에서 오류를 제한할 수 있으며 바로 수정 할 수 있다는 장점이 있다.

○ OMR, OCR/ICR에서는 현지에서 바로 내검하는 것이 불가능하다.

❑ 수작업 내검과 자동내검

○ 수작업 내검은 일련의 조사과정에서 이루어지는데 이는 조사원, 관리자, 현장 지도자, 코딩자, 입력원 등에 의해 이루어진다.

수작업의 단점은 매우 많은 시간이 얼리며 노동력이 많이 소모되며 비용이 많이 든다는 것이다.

만약 자료양이 적다면 수작업 내검만으로도 충분할 것이다.

○ 자동내검은 시간을 줄여주며 인간이 범할 수 있는 오류를 줄여주며 내검절차를 명확히 할 수 있어서 다시 처음으로 되돌리는 것도 가능하다.

○ 수작업 내검과는 달리 자동내검은 편리하며 조사표 내의 다른 정보를 기준으로 대체할 수 있다.

❑ 과도한 내검의 폐해

○ 시의성을 저해

○ 비용을 증가시킴

○ 참값을 왜곡시킬 수 있음

○ 보안에 대해 문제를 일으킬 수 있음

❑ 다른 고려사항

○ 오류를 찾는데 있어서 허용치의 결정
센서스의 모든 항목에 있어서 어떤 소수의 사람들은 허용할만한 응답을 주지 않는다.
모든 잘못된 항목을 고치려 할 수는 없으며 따라서 고칠만한 가치가 있는 항목을 찾아야한다.

○ 허용치란 (Tolerance level) 유효하지 않고 일관성이 없는 응답 중에서 내검팀에서 사전에 허용하겠다고 합의한 숫자를 의미한다.

○ 핵심이 되는 항목(Key item) 예를 들면 **연령**이나 **성별** 등은 일반적으로 허용치가 1~2%로 낮지만 덜 중요한 항목은 예를 들면 문맹률이나 장애등 은 5~10%로 높다.

○ 수정은 조사원들이 현장으로 되돌아가 다시 조사하거나 전화로 다시 인터뷰하거나 그 지역의 다른 전문적 자료를 이용하게 된다.

○ 내검 과정에서 배울 점
긍정적이나 부정적인 피드백 고리는 현재와 미래 센서스에서 데이터의 질을 향상시킬 수 있다.

❑ 내검의 비용

○ 지난 20년간 내검에 드는 비용은 줄어들지 않았으나 계속적인 기술 발전에 비하면 무리한 정도는 아니다. 일반적으로 내검활동은 다른 센서스 활동에 비해 과도하게 많은 시간을 필요로 해서 인건비도 많이 든다.

○ 과도한 내검은 센서스 결과공표를 지연시킬 수 있다.

❑ 자료저장

내검이 되고 되지 않은 자료는 모두 나중에 분석을 위해서 보관되어야 한다.

9. Session 9 : Recommendations & Conclusions

(1) 자료입력의 방법과 관련해서 각 국가는 다음 센서스를 준비하는데 특정 기술을 도입하는데 있어서 그들의 능력이나 비용을 고려해야 한다. OMR, OCR, ICR , PDA, 인터넷 조사 방법 등 다양한 방법이 논의되었으나 가장 적정한 방법이라는 것은 없으며 최신 트렌드를 따라가는 것이 아니라 각 국가의 요구에 의해 결정되어야 한다.

(2) 각 국가의 통계청은 직원들이 관계지식을 컨설턴트나 솔루션 제공자에게 얻을 수 있도록 지원해 주어야 한다.

(3) 회의에서 자료입력의 외주화 방안에 대해 논의하였는데 참가자들은 외주는 테스트와 특별한 기술방식을 실험해 보기에 충분한 시간을 두고 결정되어야 하며 다음과 같은 사항들이 고려되어야 한다.

① 입찰이 되기 전에 충분한 기술평가가 필요하다
② 계약은 엄격한 정보보호와 보안이 요구되며 품질관리계획도 포함되어야 한다.
③ 전 세계적인 계약표준이 만들어져야 한다.

(4) 자료입력의 전 과정에서 품질 보증계획이 있어야 한다. 다양한 과정이 적지만 믿을만한 지표로 모니터 되어야 하며 정상적인 것을 잘못 인식하는 오류에 특별한 관심을 기울여야 하며 인식하지 못하는 경우도 마찬가지이다.

(5) 자료입력에 대한 사전준비의 중요성과 시험조사에서 자료입력과 자료 내검이 반드시 포함되어야 한다.

(6) 통계청은 내검과 자유기입식 질문의 코드화를 위해 그 분야의 전문가들과 공조해야한다.

(7) 통계청은 통계적으로 안전한 내검과 무응답대체 전략을 수립해야 한다. 과다 내검을 피해야하지만 어떤 내검 방식도 모든 오류를 규명할 수는 없다. 따라서 통계청은 센서스를 준비하는 모든 과정에서 즉 조사표 디자인과 조사원 훈련 등 전 분야에 걸쳐 특별한 관심을 기울여야 할 것이다.

(8) 통계청은 수행된 센서스자료의 가치와 품질과 사용가치를 향상시키기 위해서 국제적인 기준에 의한 데이터자료화, 아카이브 등이 필요하다. 예를 들어 마이크로소프트사의 Microdata Management Toolkit이 있다.

(9) 통계청은 적절한 국제. 내적 법규화를 통해서 연구자에 의한 마이크로데이터의 안전한 접근이 가능하도록 해야 한다. 이는 정책수립과 자료에 입각한 의사결정을 위해 필요하다.

(10) 워크숍에서 참가자들은 그들이 관련된 경험을 공유할 수 있었고 센서스 자료처리과정을 향상시키기 위해서 자료수집과 배포과정 예를 들면 내검 프로그램 등을 2010 센서스 프로그램 웹사이트를 통해 제공해 주기를 요구한다.

(11) 통계청은 경험을 공유하는 것과 기술적인 동업이나 연구 투어등 센서스 프로세스 전 과정에 걸쳐 최선의 연습을 할 수 있도록 권장해야 한다.

Ⅳ 각 국의 자료처리 경험 토의, 공유

1. 인도

   16개 언어로 조사표 작성, 조사원 200만 명이며 자료처리에 3~5년 걸림 15개 지역에서 분산하여 처리함

2. PDA 이용 예정 국가들
   오만, 콜롬비아, 브라질 등 라틴아메리카 국가들

3. 탄자니아에서 조사원에게 유니폼 지급하여 홍보효과 있었음
   수단에서는 조사표가 자료처리 도중 화재로 소각됨->백업시스템 필요함

4. 탄자니아 조사표에서는 지문(finger print)을 찍는 난이 있는데 UNSD에서는 지문이나 바코드 등을 조사표에서 사용하는 것은 개인 식별이 가능하므로 이용하지 않는 것이 좋다고 조언함
   센서스 조사목적은 개인정보 수집이 아니라 통계작성임을 기억할 것

5. 필리핀에서는 2000년 센서스에서 ICR을 위한 깨끗하고 읽기 쉬운 조사표를 고수하였음.
   센서스 도중에 조사표가 부족하여 급하게 다시 찍었는데 원래 조사표와는 다르게 인쇄되어서 자료입력과 결과공표가 지연되었다.

6. 프랑스에서는 산업부문에서 70%는 자동코딩이 되고 10%는 코딩이 되지 못했다. 자동코딩기계가 코딩하지 못할 때 수작업으로 작업해야함
   직업 부문에서는 50%가 자동 코딩됨.

7. 캄보디아에서는 더블체크를 위해 자료입력을 두 번하였음