통계자료의 특성과 비밀보호 방법에 관한 연구

2004. 12.

<마크> 통 계 청

제 출 문

본 보고서를 2004년도 연구과제인「통계자료의 특성 및 비밀보호 방법에 관한 연구」의 연구결과 보고서로 제출합니다.

2004년 12월

통계연구과장 최 봉 호

연 구 자 : 박 원환
황 조연

요 약 문

과 제 명	통계자료의 특성과 비밀보호 방법에 관한 연구		
중심단어	마이크로 데이터, 다차원 구조, 비밀보호, 실명화, 익명화		
연구기관	통계기획국 통계연구과	연 구 자	박원환, 황조연
연구기간	2004. 6. ~ 2004. 7.(2개월)		

통계자료는 일반적으로 매크로 데이터(macro data)와 마이크로 데이터(micro data)로 구분한다. 매크로 데이터는 수집된 자료를 집계 및 요약한 자료를 말하며, 마이크로 데이터는 수집 자료에 포함되어 있는 올바르지 못한 내용을 수정한 유효자료(valid data)를 말한다.

통계자료에는 개인 및 기업의 민감한 개별정보가 포함되어 있을 수 있으므로 이러한 정보는 관련법에서 보호하도록 규정하고 있다. 그러므로 매크로 및 마이크로 데이터를 통하여 개별정보가 외부에 노출되는 일이 없도록 하여야 하는 것이 통계작성기관의 의무라 할 수 있다.

따라서 본 논문에서 통계자료의 이용 활성화와 응답자 보호를 위하여 통계자료 자체의 특성을 구조, 절차 등의 측면에서 분석한다. 그리고 매크로 및 마이크로 데이터를 이용자에 제공할 때 발생할 수 있는 비밀노출 위험(disclosure risk)을 분석하고, 이를 통제할 수 있는 방법을 제시한다. 매크로 데이터에 대해서는 빈도에 따른 비밀노출 위험에 대하여 중점적으로 알아보고, 마이크로 데이터는 실명화를 방지하기 위한 마스킹(masking) 방법에 대하여 연구한다. 뿐만 아니라 통계청의 통계자료 서비스 현황분석과 비밀노출 위험을 분석하여 연구한 보안방법을 활용하여 보완하는 방안을 제안한다.

목 차

제1장 서 론 ····· 1

제2장 통계자료의 특성 ····· 3
제1절 통계자료의 분류 ····· 3
제2절 통계자료의 특성 ····· 5
제3절 통계자료 특성분석 결과 ····· 17

제3장 통계자료 서비스 유형과 비밀노출 위험 ····· 19
제1절 통계자료 이용자와 서비스 유형 ····· 19
제2절 마이크로 데이터에 대한 공급자와 수요자의 관점 ····· 22
제3절 통계자료 이용 활성화를 위한 제도와 비밀보호 ····· 27

제4장 통계자료의 비밀보호를 위한 보안기술 ····· 30
제1절 보안침해 유형과 일반적 보안대책 ····· 30
제2절 데이터 유형별 보안위험 분석 ····· 33
제3절 비밀노출 방지를 위한 보안기술 ····· 41

제5장 비밀보호 현황 및 개선방안 ····· 58
제1절 마이크로 데이터 서비스 제도 ····· 58
제2절 마이크로 데이터 서비스 범위 ····· 63
제3절 비밀노출 방지 현황 ····· 66
제4절 비밀노출 방지를 위한 방법 제안 ····· 71

제6장 결론 및 향후과제 ····· 77

표 목 차

<표1> 통계청의 마이크로 데이터 주문통계서비스 건수 ····· 6
<표2> 통계자료의 예 : 경제활동인구조사 ····· 9
<표3> 통계자료의 3가지 속성 ····· 12
<표4> 마이크로 데이터 서비스 유형과 보안위험 ····· 20
<표5> 마이크로 데이터의 서비스 유형별 보안위험 ····· 33
<표6> 익명화 되지 않은 자료의 비밀노출 위험도 분석표 ····· 36
<표7> 익명화 자료의 보안위험 분석표 ····· 38
<표8> 데이터 링케이지 예 ····· 39
<표9> 마이크로 데이터 서비스를 위한 비밀노출 방지대책 종합 ····· 40
<표10>교육정도 통계표(가상의 통계표) ····· 41
<표11>교육정도 통계표(비밀노출 셀에 대하여 감추기 작업결과) ····· 42
<표12>교육정도 통계표(비밀노출 셀에 대하여 감추기 셀 조정) ····· 43
<표13>교육정도 통계표(랜덤 올림방법 적용) ····· 44
<표14>교육정도 통계표(올림방법 적용) ····· 45
<표15>가상 마이크로 데이터 : Alpha 카운티의 모든 레코드 ····· 46
<표16>가상 마이크로 데이터 : 일부를 교환 편집한 결과 ····· 47
<표17>교육정도 통계표(원자료 편집방법 적용) ····· 47
<표18>가상 마이크로 데이터 : 표본, 식별자 제거 등 적용 ····· 50
<표19>가상 마이크로 데이터 : 소득구분, 상한, 하한 구간조정 ····· 51
<표20>가상 마이크로 데이터 : 잡음추가 방법 적용 예 ····· 52
<표21>가상 마이크로 데이터 : 공백과 대체방법 ····· 53
<표22>통계자료 제공 규정에 정의한 자료유형 ····· 59
<표23>통계청의 마이크로 데이터 서비스 제도 및 내용 ····· 59
<표24>주문통계 서비스 현황 ····· 61
<표25>마이크로 데이터 CD-ROM 판매현황 ····· 62
<표26>통계청 통계작성 현황 : 52종 ····· 63
<표27>이용도가 높은 마이크로 데이터 종류 및 특징 ····· 64
<표28>매크로 데이터의 비밀노출 방지 방법 ····· 67

그 림 목 차

[그림1] Survey Cycle and Data Transformation ····· 4
[그림2] 통계자료의 논리적 구조 : Data Cube ····· 7
[그림3] 다차원 구조의 구성 요소별 계층구조 예 ····· 8
[그림4] 통계자료 변환과정과 메타 데이터의 관계 ····· 10
[그림5] 업무 절차상의 메타 데이터 ····· 13
[그림6] 메타 데이터의 통합방법 ····· 15
[그림7] 통계공급자와 수요자의 관점 관계 ····· 25
[그림8] 통계자료에 대한 일반적 보안대책 ····· 32
[그림9] 자료유형과 서비스 제도와의 관계 ····· 35
[그림10] 익명화 기술적용에 따른 효과분석 ····· 49
[그림11] 워터마킹 개념도 ····· 54
[그림12] 마이크로 데이터 불법복제 방지 개념 ····· 56
[그림13] 이용자 및 불법사용 관리 절차 ····· 57
[그림14] 통계자료제공 처리절차 ····· 61
[그림15] 데이터 링케이지를 이용한 개별정보 노출의 예 ····· 73
[그림16] 마이크로 데이터의 불법복제 방지 방법 ····· 79
[그림17] 불량질의 및 응답 관리 방법 ····· 73

제 1 장 서 론

이미 우리가 잘 알고 있는 것처럼 정보화 사회란 정보 그 자체의 중요성이 엄청나게 증대하고, 이를 바탕으로 하여 정보의 생산, 유통 및 이용이 기존 사회를 새롭게 바꾸는 그러한 사회를 말한다. 결국 정보화 사회에서는 정보가 유일하거나 가장 중요한 부의 원천이자 권력의 중심에 있다. 즉 정보 그 자체가 가치를 가질뿐더러 또 다른 새로운 가치를 만들어 내는 사회, 그래서 정보화 내지는 컴퓨터화를 통하여 근본적인 변화를 받고 있는 사회가 바로 정보화 사회이다.

그런데 정보화 사회에서 사생활 자유가 더 침해될 가능성이 높은지는 정보통신기술의 발전으로 전달과 복제 방법이 단순해지고 접근성이 높아진 것이 가장 큰 요인이라 할 수 있다. 게다가 이러한 개인의 사생활 및 개인정보의 침해는 한 국가의 물리적인 경계로만 한정되지 않는 특징을 갖고 있다. 따라서 정보화 사회에서 개인정보의 보호는 더 이상 한 국가만의 문제가 아닌 전 세계적인 문제로 되고 있다. 특히 이러한 정보기술의 발전으로 인하여 이제 우리는 어디에서든지 컴퓨터에 접속하여 네트워크에 연결될 수 있는 환경, 즉 '유비쿼터스' 시대로 진화하는 현실 속에 있다. 유비쿼터스 환경은 다양한 컴퓨터가 다양한 환경에서 상호 연결되는 것을 기본으로 하는 바, 이는 차세대 IT혁명으로서의 사회·경제적 변혁의 총체라 할 수 있다. 하지만, 유비쿼터스라는 새로운 IT혁명으로 누릴 수 있는 삶의 풍요로움 뒤에는 우리가 앞으로 해결해야 할 개인정보보호라는 문제가 도사리고 있다.

이와 같이 정보화의 진전과 그에 따른 역기능은 통계자료에도 예외는 아니다. 그 이유는 가구 또는 개별 사업장의 응답자로부터 취득한 정보가 개인의 정보에 해당하기 때문이며, 또한 이를 집계 분석하여 국가정책 및 국민의 의사결정을 기초 자료로 제공하고 있기 때문이다. 정보화 사회에서 통계

자료의 보급 및 활용수단은 당연히 네트워크 등 정보기술을 활용하고 있다. 이용자에게 보급이 가능한 통계자료[1]의 범위는 어디까지이고, 개별정보의 식별 위험성은 어떠한 것이 있으며, 방지를 위한 대책은 어떠한 것이 있는지 등이 필요하다. 통계자료 공급자와 이용자는 관점의 차이가 있는데 공급자는 비밀노출 위험성을 최소화하기 위하여 공급 내용을 최소화 하려고 하고, 이용자는 가치가 높은 자료를 이용하려고 한다. 그 결과 공급자와 이용자의 관계를 긴장관계에 있다고 볼 수 있다. 또한 과거에는 대형 컴퓨터를 구비하고 있는 공급자만 대량의 데이터 분석이 가능하였던 작업을 최근 정보기술의 발전으로 통계 이용자들이 직접 개인용 컴퓨터를 활용하여 집계가 가능한 환경으로 변화함으로써 데이터 량이 많은 마이크로 데이터의 요구가 증대하고 있다. 이와 같이 여건 변화하였고, 또한 향후에도 더욱 더 이러한 여건변화는 심화될 것이므로 데이터 활용범위 확대에 필요한 비밀노출 방지방법에 대하여 미국, UN, EUROSTAT 등에서 최근에 많은 연구를 하고 있다.[1][2][3][4].

통계자료는 기획, 현장조사, 집계 및 분석, 제공 등의 절차를 거쳐 작성되며, 집계자료를 매크로(macro) 데이터라 하며, 집계하지 않은 개별 자료를 마이크로(micro) 데이터라 한다. 개인정보 노출 위험(disclosure risk)은 두 가지 유형에서 모두 존재한다. 그러므로 본 논문에서는 통계자료의 특성 분석, 매크로 및 마이크로 데이터의 비밀노출 위험도를 분석하여 그에 대한 방지방법에 대하여 알아본다.

본 논문의 2장에서 통계자료의 종류, 특성에 대하여 알아보고, 3장에서는 마이크로 데이터의 공급자와 수요자의 관점, 통계자료 활용 활성화를 위한 제도와 일반적인 비밀보호 정책에 대하여 알아본다. 4장에서는 통계자료의 비밀보호를 위한 보안기술, 5장에서는 통계청의 마이크로 데이터의 서비스 현황을 분석하여 개선방안을 제안하고, 마지막으로 결론 및 향후과제를 밝힌다.

---

1) 일반적으로 통계자료는 매크로(macro) 데이터와 마이크로(micro) 데이터 모두를 말함

제 2 장 통계자료의 특성

제 1절 통계자료의 분류

통계자료[2]는 여러 가지 방법으로 분류할 수 있지만, 여기서는 분야별, 조사대상, 가공절차에 의한 분류, 취급 자료에 따른 분류 등으로 분류한다[5].

1. 분야별 분류

자료의 내용에 따라 분류하는 방법으로 사회통계, 인구통계, 산업통계, 서비스업 통계 등과 같이 분류하는 것이다.

2. 조사 대상에 의한 분류

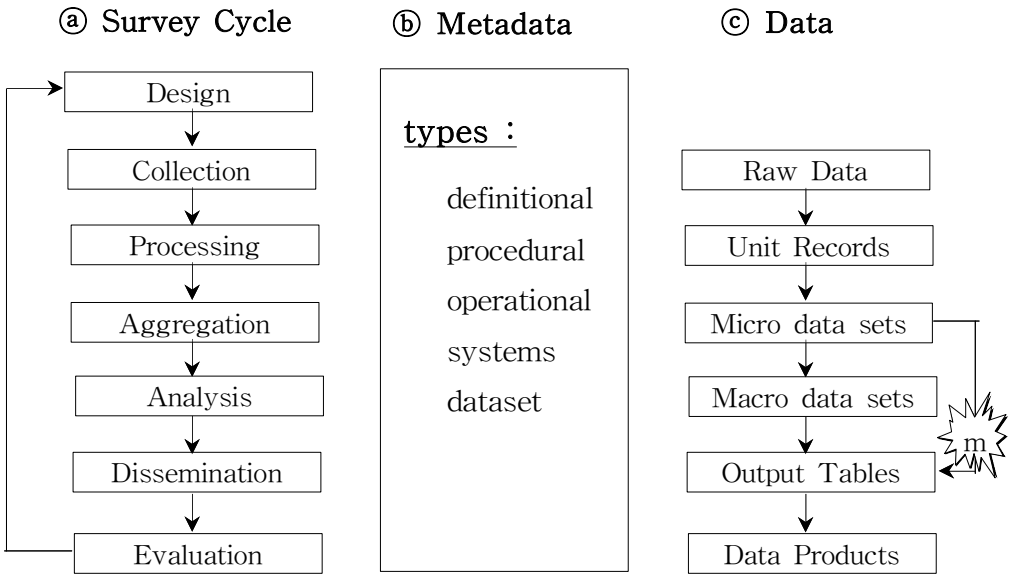
조사통계의 경우 조사대상이 전수를 대상하는 조사 자료와 표본을 대상하여 조사한 자료로 구분하는 방법이다.

3. 조사의 업무절차에 의한 분류

조사된 원시자료(raw data), 조사표 입력 레코드(unit records), 마이크로 데이터(micro data), 집계한 매크로 데이터(macro data), 통계표 자료(out tables) 등으로 분류하는 것으로 [그림1]과 같다[6]. 이 방법은 통계조사 생명 사이클(survey life cycle)의 순서에 따라 자료들을 구분하며, 이들 단계는 일반적으로 통계의 조사표 설계(design), 현장조사 및 취합(collection), 집계(aggregation), 분석(analysis), 공표(dissemination), 결과의 평가(evaluation) 등으로 나눈다. 그러나 조사단계에 의한 통계자료의 분류는 자료보다는 업무절차에 중점을 두고 있으므로 엄격한 의미에서는 통계자료의 분류라 볼 수 없다. 예를 들어 [그림1]에서 현장조사 및 취합단계(collection)와 대응하는 원시자료(raw data)가 통계자료에 해당한다.

---

2) 여기서 취급하는 자료는 조사통계자료에 한한다.



[그림1] Survey Cycle and Data Transformation

4. 자료의 가공절차(또는 변환절차)에 의한 분류

[그림1]의 "ⓒ data"의 자료 분류, 즉 원시자료(raw data), 단위 레코드(unit records), 마이크로 데이터(micro data), 매크로 데이터(macro data), 통계표(tables), 최종 산출물(data products) 등은 자료를 가공절차에 따라 분류하는 자료 형태이다. 각 자료들의 수록매체는 종이, 자기 테이프, 디스크 저장장치, CD-ROM, 보고서 등 통계자료를 수록할 수 있는 모든 것이 매체들이 해당한다.

원시자료는 응답자(respondents)로부터 얻어지는 자료, 또는 행정자료, 다른 원시자료로부터 변경된 자료가 될 수 있지만, 본 논문에서는 조사통계 자료를 중심으로 하므로 응답자로부터 획득한 자료라고 정의한다. 이 원시자료에는 응답오류, 입력오류, 무응답 등 오류들이 포함되어 있다. 이러한 오류의 자료를 기본단위 자료(unit record)별로 에디팅[3](editing), 대체방법(imputation)

3) 에디팅(editing)은 조사단계 또는 입력단계에서의 오류자료를 바로잡는 작업을 말함

등을 통하여 오류 없는 완벽한 자료로 변환하면 마이크로 데이터가 된다. 이 마이크로 데이터를 집계(aggregation), 추정(estimation) 또는 가중화(weighting) 과정을 거치면 매크로 데이터로 변환된다. 이 매크로 데이터를 임의의 항목, 지역 등을 선택(selection)하여 정해진 형식으로 변환(formatting)하여 통계표(tables)가 만들어 진다. 마지막 단계로 이 통계표에 자료를 설명하는 주석과 분석적인 설명 등을 첨가하여 최종 통계 산출물이 완성된다. 이와 같이 가공절차에 의한 자료유형은 1 사이클의 경우이며, 수차례 반복함으로써 통계자료의 특성인 시계열(time series) 형태의 자료로 자연스럽게 변화된다.

제 2절 통계자료의 특성

1. 가공절차에 따른 통계자료의 특성

[그림1]에서 눈여겨 볼 부분은 마이크로 데이터에서 결과표로 직접 변환하는 부분인 "ⓒ data"의 “m"부분은 두 가지 측면에서 특징이 있다.

첫째 특징은 최근 정보기술의 발전으로 전산장비의 성능이 향상되어 대량의 자료를 직접 최종결과로 만들 수 있는 전산장비가 출현하였기 때문에 이러한 작업이 가능해 졌다는 점이다. 과거에는 대량의 자료를 직접 취급할 수 있는 컴퓨터가 없었거나, 업무추진의 효율성이 떨어지는 문제가 있어 중간단계로 매크로 자료를 먼저 생성한 후, 이를 재작업하여 최종 통계표를 생산하는 과정을 거쳤다.

두 번째 특징으로는 통계자료 이용자들의 성향이 보다 세분화된 자료에 대한 요구로 변화하고 있다는 점이다. 과거에는 주로 보고서 형태의 통계자료의 이용자들이 주를 이루었으나, 최근 들어 다양한 형태의 자료를 요구하고 있을 뿐만 아니라 컴퓨터와 프로그램을 이용하여 직접 분석하기 위하여 마이크로 데이터에 대한 수요가 증가하고 있다. 이는 첫째 특징에서 언급하였던 바와 같이 컴퓨터의 성능의 향상과 가격의 하락으로 개인이 직접 컴퓨터의 활용이 가능한 환경으로 변화하였다는 점도 그 원인이라 할 수 있다.

<표1> 통계청의 마이크로 데이터 주문통계서비스 건수[7]

년도	2000년	2001년	2002년	2003년
건수	441	514	559	536

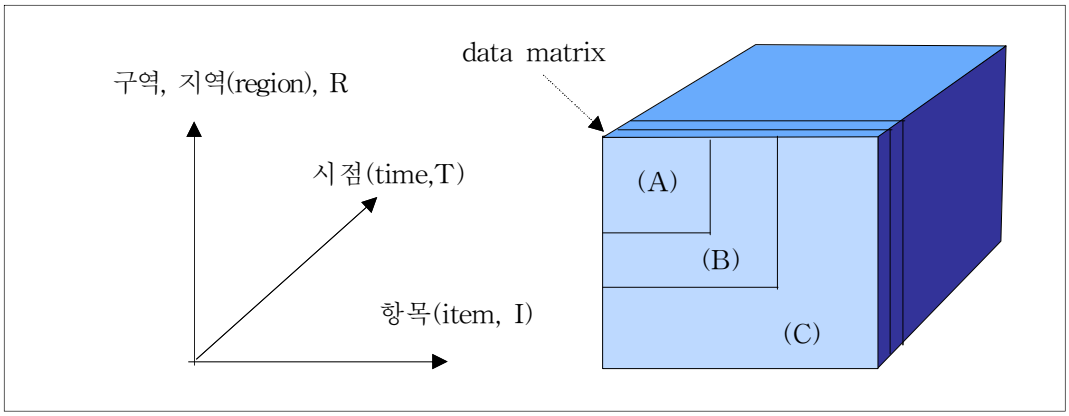
이상의 두 가지 특징은 정보화 사회로 이행과 사회의 복잡· 다양화 추세의 영향으로 볼 수도 있지만, 과거에는 전산기의 자료처리 능력이 부족하여 분석이 불가능한 점이 있었으나 최근 발전된 전산장비의 성능이 통계자료의 다양한 분석이 가능해진 점이 주요 요인을 보인다. 따라서 향후에도 지속적으로 전산장비의 성능은 향상될 것이므로 통계자료 수요 또한 점점 더 세분화 되고 다양해질 것으로 예측되어 진다.

이상과 같이 통계자료의 이용 방법 및 이용자 요구 성향의 변화는 통계자료를 생산하고 공급하는 통계기관의 입장에서는 또 다른 과제를 갖게 되었다. 즉 개인정보보호에 관한 사항이다. 물론 통계자료의 경우 통계법을 우선 적용하도록 되어 있어 개인정보보호법과는 직접 관련이 없다고 하지만, 통계법에는 조사응답자에 비밀보호 사항(14조)와 위반시의 벌칙사항(23조)을 규정하고 있어 당연히 보호하여야 한다. 반면에 통계법에는 또한 통계자료의 활용(16조)을 널리 하도록 하고 있다. 그 결과 이용자의 세분된 자료의 요구와 응답자 정보보호라는 상반된 두 가지 과제의 해결이 큰 숙제로 남는다. 이 과제의 해결을 위하여 노출통제(disclosure limitation)라는 연구 분야가 있다. 노출통제 또는 비밀노출 방지 문제는 뒤에서 취급하가로 하고, 우선 가공절차 상의 통계자료들에 대한 구조적 특성에 대하여 알아본다.

2. 통계자료의 구조적 특성 : 다차원 구조

통계자료 이용자는 [그림1]의 ⓒ 자료를 활용하고자 하므로 이들 자료의 구조적 특성을 분석하고자 한다. 이를 위하여 통계자료를 논리적 구조로 구조화 하면 [그림2]와 같다. 통계조사의 1사이클, 즉 [그림1]과 1 주기가 진행된 자료는 Data Matrix로 표현되는 2차원 구조이며, 통계조사 사이클이 수차례 쌓

이면 다차원 구조로 만들어 지면서 시계열 형태의 자료로 표현된다. 이를 데이터 큐브(data cube)라고도 한다.



[그림2] 통계자료의 논리적 구조 : Data Cube

[그림2]는 이와 같은 다차원 구조와 이를 이루고 있는 구성요소들을 나타내고 있다. 통계자료는 1회에 조사한 경우에는 기본적으로 조사한 자료(A), 대체한 자료(B, imputed data), 그리고 집계자료(C, aggregated data)로 이루어진다. 하지만, 수차례의 반복 조사를 통하여 이들 자료들은 3차원 이상으로 변화하게 되어 다차원 구조를 이룬다. 3차원 이상이라 함은 다차원의 구성요소인 지역(R),시점(T), 항목(I) 등이 [그림3]과 같이 각각 계층구조를 이루고 있기 때문이다.

[그림2]에서 (A)는 실제 조사한 자료로써 표본조사인 경우는 (A) 자료를 기초로 (B)를 대체 또는 확장한 자료이며, 전수조사의 경우는 레코드의 증가가 없으므로 (A)와 (B)가 통합한 자료가 조사한 자료이지만, 항목 간에는 오류자료 수정을 위하여 대체법(imputation)을 적용하기도 한다. (C)자료는 [그림3]과 같이 다차원 구조의 구성요소별 계층구조 중 필요한 단계별로 집계하여 변환된 자료이다.

이들 (A), (B), (C) 자료의 양은 실제적으로 조사된 자료가 많다고 생각하지만, 실은 [그림3]의 계층구조별 모든 집계를 한다면, (C)이 가장 많다. 그 이유는

항목의 각 단계, 지역의 각 단계, 시점의 각 단계에 대한 모든 경우의 수에 대하여 집계를 한다면, 엄청난 양의 자료가 만들어 진다. 하지만, 실제로는 자료 자체의 유용성 부족, 보고서 지면의 한계 등 여러 가지 이유로 특정한 일부만을 집계하고, 분석, 활용하고 또한 서비스하고 있다. 그 결과 집계하지 않고 통계자료 이용자에게 서비스 하지 않고 있는 특정한 자료에 대하여 별도의 주문[4]에 따라 서비스하기도 한다.

◆ 항목(Item) : 대분류 - 중분류 - 소분류 - 세분류 - 세세분류 - 품목분류
◆ 지역(Region) : 전국 - 광역시도 - 구시군 - 읍면동 - 조사구
◆ 시점(Time) : 10년 - 5년 - 년 - 반기 - 분기 - 월 - 시

[그림3] 다차원 구조의 구성요소별 계층구조(예)

한편, 이러한 계층적 구조의 특성으로 인하여 통계자료를 데이터베이스에 수록할 경우 다단계 접근경로로 구현할 수 밖에 없는 특성을 가진다. 그 결과 통계정보시스템(KOSIS)[5]과 같은 정보시스템의 이용자는 접근이 복잡하고, 불편한 단점이 있다. 그러므로 이와 같은 정보시스템의 통계데이터베이스를 설계할 때, 이러한 점을 고려하여 설계하여야 한다.

3. 통계자료의 속성별 특성 : 범주속성, 요약속성, 메타 데이터

통계자료 외형을 보면 <표2>와 같이 통계표 형식으로 되어 있다. <표2>는 경제활동인구조사의 여러 통계표 중의 일부로써 전국의 경제활동인구조사 중 15세 이상 인구, 경제활동인구, 취업자, 실업자 등을 연령계층별로 나타내고 있다. 이 표를 자세히 살펴보면 구성형태가 세 가지 유형의 자료, 즉 표두와 표측 자료, 통계수치자료, 주석 및 단위 자료 등으로 구성되어 있음을 알 수 있다.

4) 통계청의 경우 “주문통계서비스” 제도가 그 예임
5) 통계청 통계정보시스템(KOSIS : KOrea Statistical Information System)을 말함

이들 자료 중에서 표두와 표측 자료는 앞에서 논의하였던 다차원 구조에서의 품목, 지역, 시점의 조합으로 구성된 자료로써 통계수치자료의 범위를 정하는 특성을 지니고 있다. 그러므로 이 자료를 범주속성 자료(category attribute data)라고 한다. 이에 반하여 통계수치자료를 요약속성자료(summary attribute data)라고 한다. 또한 주석 및 단위 등의 자료는 자료의 설명을 위한 자료, 또는 자료의 성격을 기술하는 자료는 메타 데이터(meta data)라 한다[5]. 통계자료에서 이들 3가지 유형의 자료 모두가 나름대로 자기의 역할을 하고 있으며, 특히 메타 데이터의 역할이 마이크로 데이터의 통계자료의 이용자에게는 상당히 중요한 비중을 차지한다.

<표2> 통계자료의 예 : 경제활동인구조사[6)]

경제활동인구조사(2004. 4월)

(단위:천명)

연령구분	15세이상 인구	경제활동인구	취업자	실업자
계	37,639	23,482	22,673	809
15-19세	3,081	251	219	32
+20-29세	7,098	4,682	4,338	344
+30-39세	8,519	6,427	6,238	188
+40-49세	8,002	6,373	6,233	140
+50-59세	4,861	3,397	3,318	79
+60세이상	6,077	2,352	2,326	26

<항목주석>
1) 15세이상 인구 중 군인, 전투경찰, 공익근무요원, 형이 확정된 교도소 수감자, 외국인 등은 제외됨

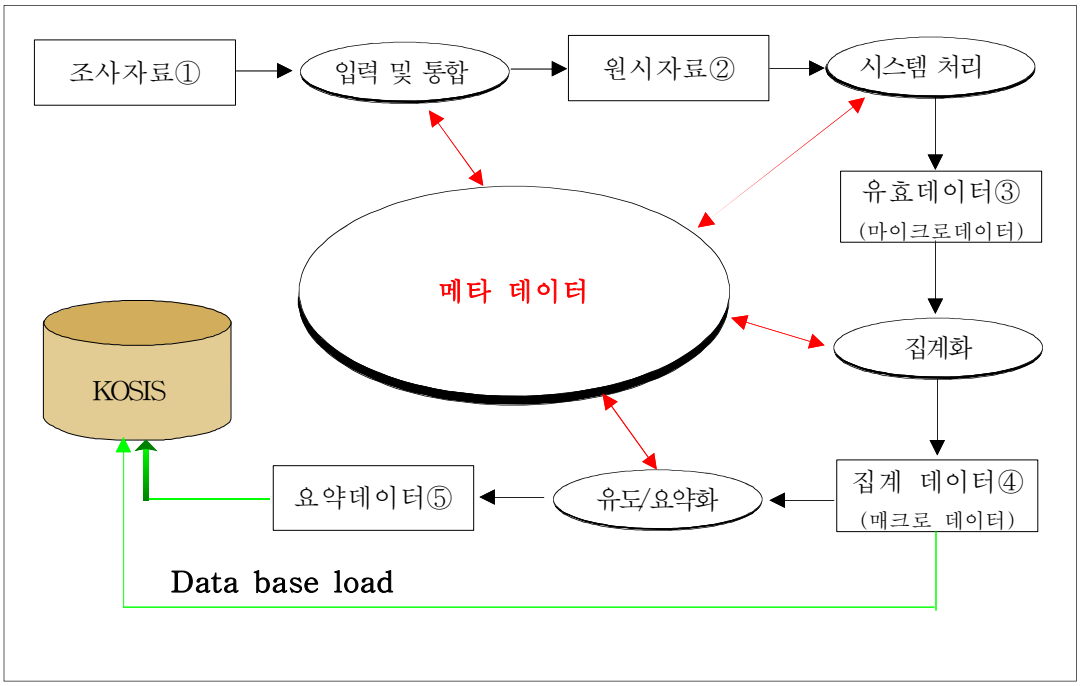
<통계표 주석>
- 분기자료의 경우 "82년 2/4분기까지는 매분기말(3,6,9,12월) 자료이며, '82년 3/4분기 이후는 분기평균자료임

- 2003년 12월에 2000년 인구주택총조사 결과를 토대로 작성된 추계인구의 변경과 연령계층별 승수의 적용으로 1991년 1월 - 2002년 12월까지의 자료가 변경되었음

[그림4]는 [그림1]에서 메타 데이터의 역할과 기능을 설명하기 위하여 변

6) 통계정보시스템(KOSIS)에 수록되어 있는 통계자료

형하여 도식화하였다. [그림4]를 살펴보면, 우선 초기에 현장에서 조사한 자료가 ①의 자료가 되며, 이는 조사표 형태에 따라 자료의 수록자료 매체가 달라진다. 종이조사표의 경우는 종이, 컴퓨터 활용조사(CAPI : Computer Assisted Personal Interviewing)의 경우 컴퓨터를 면접에 직접 이용하므로 전산입력자료 등이 될 것이다. 이와 같이 조사 자료를 전산장비에 입력하여 통합한 ②의 자료를 원시자료(raw data)라 하며, 이 자료에는 입력오류, 응답오류 등 오류자료들이 포함되어 있다. 오류의 유형은 범위오류, 논리오류, 무응답 오류 등 다양하다. 이들 오류들은 오류처리 규칙에 따라 정정하는 작업을 수행하여 바로 잡는데, 그 과정을 종합하여 “시스템 처리”라고 한다. 이 과정에서 실질적인 작업은 내용검사 및 정정, 대체법(imputation) 적용 작업을 수행한다. 그 결과는 오류 없는 유효자료인 마이크로 데이터로 전환된다. 이와 같은 집계화 과정을 거쳐 집계자료인 매크로 데이터, 유도 및 요약화 과정을 거쳐 요약 데이터가 만들어 진다. 요약 데이터는 각종 통계표를 요약한 자료이므로 집계 데이터보다 그 범위는 넓지만, 자료의 양은 그리 많지 않을 수 있다.



[그림4] 통계자료 변환과정과 메타 데이터 관계

[그림4]에서 보는 바와 같이 이러한 자료의 변화과정에서 메타 데이터는 각 과정별로 필수적으로 참조하거나 사용하고 있지만, 일반적으로 관리를 소홀히 하는 경향이 있는 데이터이다. 메타 데이터는 자료의 형식(types)에 관한 사항, 항목 등의 정의에 관한 사항, 업무절차에 관한 사항, 연산에 관한 사항, 전산시스템과 관련되는 사항, 그리고 자료의 구조(dataset)에 관한 사항 등을 말한다. 이들 메타 데이터들은 한시적으로 필요한 경우도 있지만, 통계자료의 용도가 폐기될 때까지 지속적으로 관리되면서 활용되어야 하는 자료로써 산업분류코드 목록, 행정구역코드 목록 등이 있다.

통계조사 초기부터 데이터베이스에 적재하여 서비스 단계까지의 메타 데이터 예를 살펴보면 다음과 같다.

- **조사 단계**에서의 메타 데이터는 조사항목에 대한 설명자료, 코드로 변환이 필요한 항목의 코드변환규칙 등과 같은 자료이다. 하지만, 이러한 메타 데이터는 현장조사에 관한 사항이므로 자료의 활용에 필요한 메타 데이터와는 다소간의 속성 차이는 있다. 다만, 통계자료의 이용자가 참고할 수 있도록 <표2>의 항목주석과 같은 용도의 메타 데이터로 이용될 수 있는 자료이다.

- **입력 및 통합** 과정에서의 메타 데이터는 요약속성 자료의 설계도, 즉 조사표 내용을 전산자료화 하는데 필요한 파일 설계도(file layout)가 대표적이다. 이외에도 행정구역코드, 산업분류코드 등과 같은 코딩자료와 단위 변환을 위한 규칙 자료 등이 있다.

- **시스템 처리** 단계의 메타 데이터는 요약속성자료를 중심으로 범위오류 검출규칙, 논리오류 검출규칙, 일련번호 점검규칙과 같은 오류검출을 위한 규칙, 무응답에 대한 처리를 위한 대체법 적용을 위한 규칙 등이 있다.
- **집계화 단계**에서는 요역속성 데이터의 집계를 위한 규칙이나 수식, 범주속성 데이터와의 연계방법 등이 메타 데이터에 포함된다.

◦ **유도화 및 요약화** 단계에서도 집계화 단계와 마찬가지로 유도 또는 요약을 위한 규칙과 관련 속성 정보들과의 연계 규칙 등이 메타 데이터이다.

◦ **통계 데이터베이스로의 적재(load)** 단계에서의 메타 데이터는 집계 데이터에서 DB 테이블과 연계, 요약/유도 자료와 DB테이블과의 연계자료 등과 시계열 유지를 위한 자료 등이다.

<표3> 통계자료의 3가지 속성

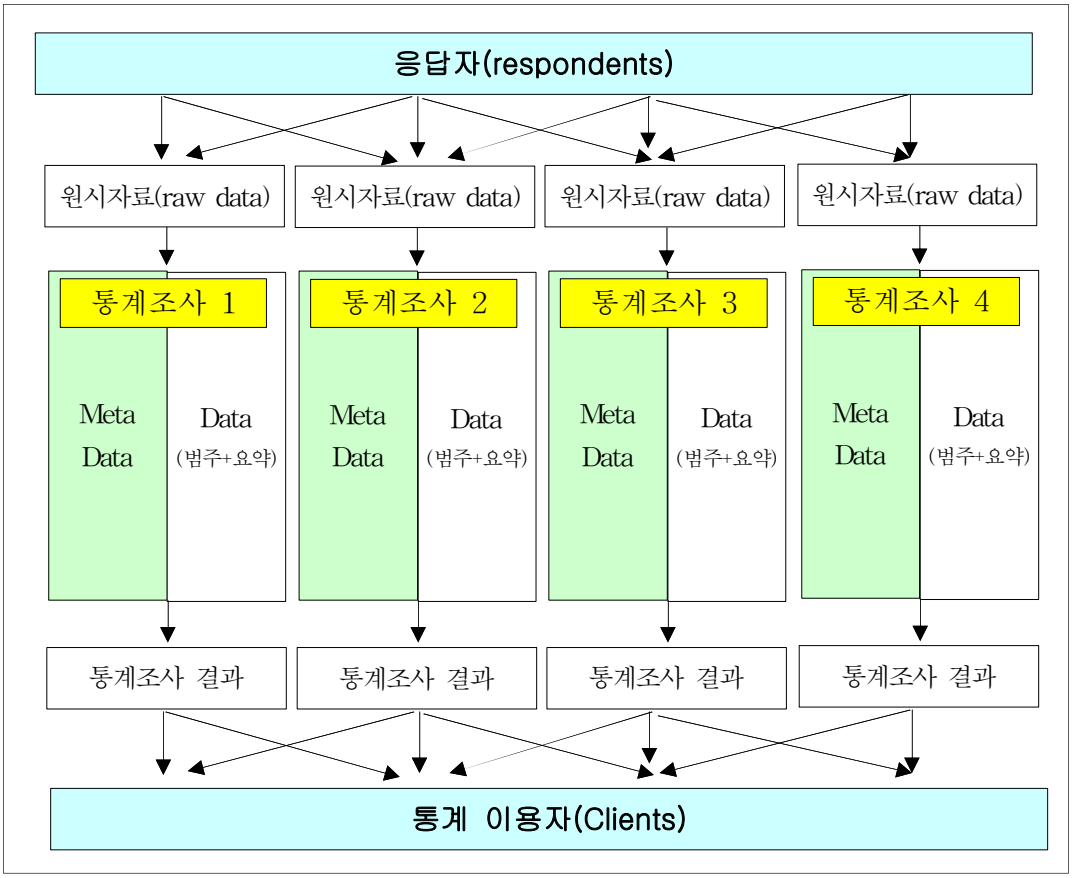
구 분		내 용	활용도 및 특징
통계자료	범주속성	◦수치자료의 범위를 지정하는 자료	◦요약속성과 결합하여 활용 ◦항목×지역×시점의 결합자료
	요약속성	◦통계수치 자료	◦단독으로 의미파악 곤란 ◦요약속성과 결합하여 활용
메타 데이터		◦통계자료의 생산규칙 등과 같은 자료를 위한 자료(data for data) ◦마이크로 데이터의 재집계에 필수적으로 필요한 자료	◦관리 소홀이 우려되는 자료 ◦전문 관리자 지정을 통하여 관리가 필요한 자료 ◦부재시에 마이크로 데이터의 활용성은 거의 없음

이상과 같이 범주속성 자료, 요약속성 자료, 메타 자료 등 3가지 유형의 속성별 특성을 살펴보았다. 이들 속성별 자료들은 단독으로 활용될 수 없을 뿐만 아니라 서로 연계되어야만 통계자료의 정확한 의미 파악이 가능할 뿐만 아니라 분석이 가능해 진다. 다시 말해서 통계자료라 함은 범주속성과 요약속성 데이터를 통합한 데이터이며, 그 의미를 보다 정확하게 전달하기 위해서는 메타 데이터의 도움이 필요하다고 할 수 있다. 반면 매크로 데이터보다 마이크로 데이터와 밀접한 관계가 있는 메타 데이터는 최초 통계조사의 결과 집계, 요약 및 유도와 관련되는 사항을 설명하는 데이터이다. 만약 이 메타 데이터의 유지관리가 부실하다면, 향후 동일 자료의 재 집계 등을 통한 재활용은 불

가능해 지거나 손실된 메타 데이터를 추적하여 확보하는데 많은 시간을 낭비해야만 한다. 따라서 통계자료의 가치를 높이고 활용을 확대하기 위해서는 요약속성 데이터를 중심으로 범주속성 데이터와 메타 데이터의 속성의 체계적인 유지관리가 필수적이라 하겠다.

4. 메타 데이터의 효율적 관리를 위한 특성분석

메터 데이터는 외형적으로 그리 중요하지 않는 것으로 보이지만 [그림5]에서 보는 바와 같이 통계조사의 업무절차(또는 자료 변환 과정)의 전 과정에 관련된 자료로써 향후 자료 재집계시에 핵심적인 자료가 된다.



[그림5] 업무 절차상의 메타 데이터

[그림5]는 통계조사의 응답자로부터 통계 이용자까지의 과정을 그림으로 표현한 것이며, [그림5]의 내부에 있는 메타 데이터와 데이터로 구성된 모듈, 즉 “통계조사1”부터 “통계조사4”까지 4가지 통계조사에 대하여 데이터 흐름(data flows)과 데이터 저장장치 간의 관계도이다. 각 조사간의 데이터(요약속성과 범주속성)와 메타 데이터가 상호 독립적으로 존재하면서 그 들의 역할을 하고 있는 것을 보이고 있다[6]. 그렇지만, 이러한 독립적인 항목에 통합적인 개념, 기능 통합, 데이터 통합 등의 부족으로 다음과 같은 4가지 문제를 야기할 수 있다.

◦ **첫째,** 응답자는 인터뷰과정에서 서로 다른 개념의 조사항목에 대하여 응답을 하여야 하고, 또한 중복적인 응답을 하여야 할 우려가 있다. 그 결과 응답자의 응답 부담은 증가하게 될 것이다.

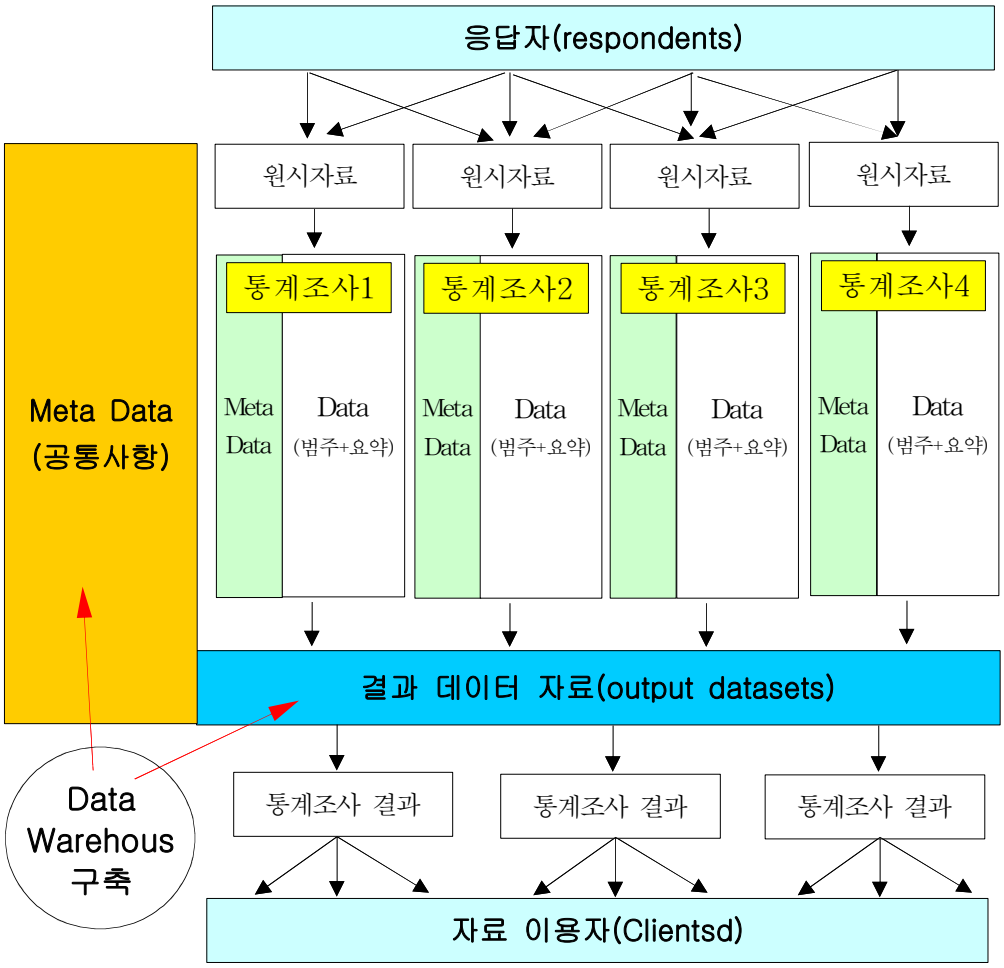
◦ **둘째,** 각 조사들 간에 공통적으로 활용할 수 있는 자료처리 부분을 중복적으로 처리해야 하고, 또한 데이터의 보급 또한 중복될 우려성이 있다. 그 결과 전산시스템의 경제적 활용이 떨어질 뿐만 아니라 시스템 자원의 유지관리에 불편을 초래할 수 있다.

◦ **셋째,** 범주속성간의 개념 차이로 인하여 통계자료의 제공을 위하여 필요한 범주속성의 항목들의 조합이 쉽지 않은 어려움이 있다.

◦ **넷째,** 통계자료 이용자는 서로 상이한 통계자료를 접하게 됨으로써 통계자료에 대한 신뢰도에 문제가 될 소지가 있다.

이상과 같은 문제점을 해결하기 위하여 통계조사 절차의 리엔지니어링과 자료관리 절차개선 등의 노력을 호주, 미국, 캐나다, 네덜란드 등에서 1970년을 전후부터 현재까지 진행 중이다[8][9][10][11]. 그 주요내용은 통계 단위(statistical units), 산업분류, 데이터 개념 등의 표준화, 에디팅과 대체법(imputation)의 일반화 등이다. 이들 분야는 대체로 메타 데이터와 관련이 있으므로 데이터의 활용 시각에서는 메타 데이터에 집결된다. 그러므로 메타 데이터에 대하여 좀더 깊이 있게 고찰해 보고자 한다.

[그림5]의 각 조사별로 독립적으로 존재하는 메타 데이터를 [그림6]과 같이 공통부분을 추출하고, 이를 표준화하여 활용하는 방법으로 전환할 경우 앞에서 언급한 4가지 문제점이 해소될 수 있다. 하지만 이를 위한 사전작업은 단순하지 않기 때문에 어려움이 많다. 그러나 장기적으로 [그림6]과 같이 메타 데이터를 종합하여 표준화해야 할 것이다.



[그림6] 메타 데이터 통합방법

[그림6]은 [그림5]에서 각 통계조사별 메타 데이터를 통합하여 공통적인 부분은 별도로 통합하고, 각 통계조사별 메타 데이터는 그대로 유지하도록 하고 있다. 통합된 메타 데이터는 데이터웨어하우즈를 활용하여 관리함으로써

각 통계조사마다 활용이 가능하게 하여야 한다. 마찬가지로 통계자료도 범주 속성과 요약속성을 통합하여 공통항목을 공유함으로써 중복조사를 배제할 수 있다. 이와 같은 메타 데이터 유형은 아래와 같이 5가지로 구분할 수 있다.[6]

◦ **개념 정의 메타 데이터**(definitional metadata) : 통계 단위(statistical units), 인구 또는 사업체, 분류, 자료항목(data items), 표준 질문(standard question), 통계 용어(statistical terminology) 등이 있다.

◦ **절차정의 메타 데이터**(procedural metadata) : 데이터를 모으고 처리하는 절차와 관련되는 메타 데이터를 말한다.

◦ **연산 메타 데이터**(operational metadata) : 정의된 절차의 구현과 관련되는 사항 또는 결과 집계(summarizing results)를 위한 사항으로 응답률(response rates), 에디팅 실패율(edit failure rates) 등이 있다.

◦ **시스템 메타 데이터**(system metadata) : 프로그램에 의하여 사용되는 사항으로 파일 설계서(file layout), 접근경로(acccess paths) 등이 있다.

◦ **자료관련 메타 데이터**(dataset metadata) : 데이터에 대한 설명 자료로써 이름, 수록내용 설명, 자료 항목, 자료 셀에 대한 주석 등을 말함

이상과 같은 메타 데이터는 유형별로 종합하여 관리함으로써 실제 마이크로 데이터는 물론 매크로 데이터의 유용성(data utility)을 높이는데 많은 도움이 될 것이다.

제 3절 통계자료 특성분석 결과

1. 통계자료라 함은 요약속성, 범주속성, 메타 데이터를 종합한 자료를 말하며, 이들을 종합하여 체계적인 관리가 필요한 자료이다.

통계자료는 수치를 나타내는 요약속성 부분과 이 수치자료의 범위를 나타내는 범주속성 자료로 구성된다. 또한 이들 자료에 대한 설명자료 또는 산식자료 등을 망라한 메타 자료는 조사단계에서부터 최종 산출물의 완성단계, 뿐만 아니라 최종서비스 단계까지 관련이 있는 자료이다. 그러므로 통계자료의 가치를 높이고, 이용 확대를 위해서는 요약속성, 범주속성 및 메타 데이터를 종합하여 체계적인 관리가 기반이 된다. 이들 3가지 속성자료의 관리를 위한 중점사항은 다음과 같다.

◦ **요약속성**의 경우는 매 조사마다 조사항목이 변화하므로 이를 사후에 활용이 편리하도록 일관되게 정리함은 물론 파일설계서(file layout)로 관리하여야 한다.

◦ **범주속성**의 경우도 매 분기마다 행정구역의 변화와 관련 코드의 변경, 산업분류 변경 등이 매 조사마다 일어나고 있으므로 파일설계서에 수록되어 있는 코드에 대한 정확한 자료의 관리에 중점을 두어야 한다.

◦ **메타 데이터**는 요약속성과 범주속성 모두에 관련되며, 요약속성의 경우는 파일설계서, 범주속성의 경우 코드, 조사단계의 조사항목의 설명자료 등 통계자료 활용에 필요한 설명 자료들이다. 또한 마이크로 데이터에서 매크로 데이터로 변화하는 과정에서 필요한 산식 등을 총망라한 자료이므로 사후 자료의 재활용에 핵심적인 역할을 하는 자료이므로 철저한 정리, 체계화가 필요한 자료이다.

2. 통계조사 업무의 최종단계는 조사결과 발표가 아니라 향후 활용을 위하여 필요한 자료를 종합하여 정리하는 단계이다.

통계조사는 기획, 현장조사, 전산입력 및 처리, 집계, 분석, 결과 발표 등의 순으로 1차적인 업무절차는 종료된다. 하지만, 통계자료는 미래에 내부적으

로 분석이 필요하여 재집계하거나 외부 연구자들이 자신들의 연구를 위하여 과거자료 요청이 예상되는 자료이다. 그러므로 결과발표 이후 자료에 대한 명확한 사양을 정리하여 이에 대비하여야 한다. 이 때 요약속성, 범주속성, 메타데이터 등을 종합·정리하여 활용함으로써 해당 통계자료의 유용성을 극대화 할 수 있다.

3. 통계자료는 시계열이 중요시 되는 자료

통계자료는 [그림2]의 논리적 구조를 가지고 있으며, 이의 구성요소는 항목(I), 지역(R) 그리고 시점(T) 등 3가지로 이루어져 있다. 이들 구성요소 중에서 시점은 통계자료의 시계열의 중요성을 의미한다. 따라서 시계열 유지를 위하여 다른 구성요소인 항목, 지역에 대한 비교기준도 명확하게 되어야만 시계열 유지가 용이하여 진다.

4. 통계자료는 수정과 갱신이 없는 반영구적인 자료

통계자료의 최종 조사결과 발표 이후에는 기준년도 개편작업 등과 같은 갱신 이외에는 자료의 수정을 불허하는 특성이 있다. 그러므로 최종 발표된 집계 또는 요약 자료와 이후에 동일 자료에 대한 처리작업 결과는 동일하여야 한다. 때문에 자료집계 과정에서 특별히 조건을 가지고 집계가 이루어 졌다면 그 조건을 메타 데이터에 명확하게 나타내어야 한다. 그렇지 않으면 향후 동일한 자료에 대한 동일한 처리 결과를 기대하기 어렵기 때문이다.

5. 메타 데이터의 표준화는 데이터의 유용성(utility)을 증대시킨다.

통계자료는 자료 자체만으로도 의미가 있지만, 그 보다는 집계, 요약 등을 통한 분석과 타 자료와의 비교를 통하여 한층 더 깊은 의미를 파악하여 의사결정에 활용하게 된다. 메타 데이터는 이러한 분석을 가능하게 하는데 필요한 자료라 할 수 있다. 그러므로 통계자료의 유용성과 메타 데이터는 밀접한 관련이 있다.

제 3 장 통계자료 서비스 유형과 비밀노출 위험

제 1절 통계자료 이용자와 서비스 유형

통계자료의 이용 목적은 이용주체에 따라 다소간의 차이는 있겠지만, 일반적으로 올바른 의사결정(decision making)을 위하여 주로 이용한다. 이를 위하여 필요한 통계자료는 크게 매크로 데이터와 마이크로 데이터로 나누어진다. 또한 이용방법은 간단하게 인용 또는 참고용 자료로 활용하는 경우와 심층적인 분석 작업에 활용하는 경우 등 다양하다. 본장에서는 이와 같이 다양한 이용자 유형과 필요한 통계자료 형태 등에 대하여 고찰한다.

1. 통계자료의 이용자 유형[7)]

통계자료는 통계간행물 또는 통계 데이터베이스 등에서 접근 가능한 자료를 이용하는 단순 이용자 계층, 심층 분석을 목적으로 하는 계층, 통계조사의 표본추출을 위하여 모집단 자료 확보를 위한 이용자 계층 등 세 가지 유형이 있다. 이상과 같은 통계 이용자 계층을 다음과 같이 구분하여 정의한다.

◦ **단순 이용자**는 통계간행물, 통계 데이터베이스 등 대외 공표된 매크로 자료를 주로 이용하는 통계자료 이용자를 말하며, 학생, 학자, 기업인 등 그 계층은 다양하다.

◦ **심층 이용자**는 단순 이용자이기도 하면서, 별도로 심층적인 분석을 위하여 마이크로 자료가 필요한 통계자료 이용자를 말하며, 석·박사 과정의 학생과 연구기관 연구원 등 학자들이 대부분을 차지하고 있다.

◦ **명부 이용자**는 통계조사를 위하여 필요한 표본조사 대상처 추출을 위하여 명부자료를 필요로 하는 이용자로 통계기관이 주를 이루고 있다.

---

7) 통계청 내부 이용자는 제외한 대외 서비스에 한함

2. 대외 서비스를 위한 통계자료 유형과 보안위험

대외 서비스 대상 통계자료는 형태적으로 볼 때 [그림1]과 [그림4]에서 보는 바와 같이 마이크로 데이터, 매크로 데이터, 요약데이터 등이 있다.

대외 공표된 매크로 데이터는 통계표 형식(tables)으로 대외 이용자에 통계 간행물, CD-ROM, 통계 데이터베이스 등으로 서비스 하고 있다. 공표되지 않은 매크로 데이터 및 마이크로 데이터는 다음 <표4>와 같이 이용자의 요구에 따라 형식과 내용 등을 달리하고 있다[4].

<표4> 마이크로 데이터 서비스 유형과 보안위험

서비스 형태		자료 특징 또는 내용	보안위험
통계표(tables) 형대로 가공하여 대외 서비스하는 방법		◦ 심층이용자 요구에 따라 생성된 통계표 ◦ 심층 이용자가 필요에 따라 재집계 등과 같은 작업이 가능하도록 매우 상세한 메트릭스 형태의 통계표 ◦ 이 경우에 지나치게 상세하면 마이크로 자료 수준의 비밀노출 위험성이 존재	있음
익명 마이크로 자료파일 (AMF)[8]	공공이용 파일 (public use file)	◦ 외부 사용자를 위하여 CD-ROM 등과 같은 매체에 수록한 통계자료 ◦ 사용 인가가 필요하지 않은 자료	있음 (무단복제)
	인가파일 (licensed file)	◦ 특정한 심층 이용자를 위하여 제공되며, 파일 제공 전에 법적 보증을 위한 서명이 필요한 자료	있음
원격접속제도(RAF)		◦ 컴퓨터 통신망을 통하여 심층 이용자들이 마이크로 데이터를 이용할 수 있도록 하는 제도	있음
데이터 실험실(DL)		◦ 통계기관[9]의 엄격한 심사 및 감독 하에 청사 내에서 상세한 마이크로 자료 접근 허용	있음

8) AMF : Anonymised Microdata Files, RAF : Remote Access Facilities, DL : Data Laboratories

9) 통계기관은 통계청 또는 통계작성기관을 말함

<표4>에 나타난 바와 같이 통계자료를 다양한 방법으로 이용자에게 공급할 수 밖에 없는 이유는 자료 속에 숨어 있는 개인비밀을 보호해야 하는 것이 첫째 이유이다. 이 문제는 지금까지 국내문제였으나 최근 정보기술의 발전으로 통신망을 통한 마이크로 데이터의 보급이 증가하면서 국제적인 문제로 확대되고 있다. <표4>에는 마이크로 데이터를 이용자에게 보급하는 방법 또는 수단을 5가지로 구분하고 있다. 각 방법에 대하여 응답자의 비밀보호를 위한 방법과 정도, 보안위험 요소를 살펴보면 다음과 같다.

◦ 통계표(tables)로 재집계하여 서비스하는 방법은 마이크로 데이터를 심층이용자의 요구에 따라 [그림2]와 같이 데이터 큐브(data cubes)로 알려져 있는 매트릭스 형태로 재집계하여 서비스하는 방법이다. 이 자료는 상세한 정도에 따라 비밀노출 위험성이 마이크로 데이터와 유사한 수준으로 존재할 수 있으므로 비밀보호 조치가 필요하다.

◦ 익명 마이크로 자료 파일 형식의 공공이용 파일(public use file)은 비밀보호 방법 적용 등 사전작업을 통하여 익명화한 마이크로 데이터를 CD-ROM 등과 같은 매체에 수록한 기성품이다. 이 방법은 비밀노출의 위험을 사전에 기술적으로 제거하였으므로 위험성은 거의 없다고 볼 수 있다. 다만, 비밀보호 방법에 따라 원 자료 가치를 심하게 훼손할 수 있기 때문에 비밀보호 방법과 자료의 유용성을 동시 만족할 수 있는 방법의 개발이 필요하다.

◦ 익명 마이크로 자료 파일 형식의 인가 파일(licensed files)은 비밀보호 방법을 최소화하여 자료의 유용성을 원 자료와 가장 가깝게 하여 특정한 심층 이용자에 제공하는 파일형식이다. 이 자료는 비밀노출의 위험성이 있는 자료이므로 사용자에 대해서는 법적인 안전장치와 이를 뒷받침하는 절차가 필요하다.

◦ 원격접속제도(RAFs)는 데이터 형식을 말하는 것이 아니며, 상기의 데이터를 전달하는 방법으로 컴퓨터 통신망을 통하여 심층 이용자가 마이크로 데이터를 이용할 수 있도록 하는 제도이다. 이 제도에서 취급하는 자료로는 익명 마이크로

자료파일 형식의 공공이용파일이나 인가파일을 주로 이용하게 한다. 인가파일의 경우는 본인인증을 위한 전산보안(computer security) 기술을 적용하여 데이터 통신보안이 완벽하여야 한다.

◦ 데이터 실험실(DL)은 마이크로 데이터 이용 장소를 데이터가 있는 곳으로 한정하고, 비밀노출 방지를 위하여 관리감독을 철저히 하는 제도를 말한다. 이 경우의 최종 결과인 매크로 데이터만 외부로 반출이 가능하다. 이 제도는 마이크로 데이터를 이용하는 방법이 복잡한 경우에 활용효과가 있는 제도이나, 이용자가 전산처리 능력이 있어야 가능한 방법이다. 그러나 비밀노출 위험이 존재하는 한 감독이 철저하여야 하며, 이용자의 의무사항을 숙지시키는 조치도 필요하다.

이상과 같이 통계자료의 대외 서비스 형태 및 제도에 대하여 알아보았다. 이와 같이 마이크로 데이터를 다양한 형식으로 서비스 할 수밖에 없는 근본적인 이유는 데이터 속에 존재하는 비밀사항의 관리에 대하여 통계기관과 심층이용자 또는 명부이용자 사이에 관점의 차이에서 비롯된다.

제 2절 마이크로 데이터에 대한 공급자와 수요자 관점

일반적으로 공산품의 공급자는 현행 기술로 보다 비싼 가격으로 보급하기를 바라며, 반면 수요자인 소비자는 고품질의 제품을 저가로 공급받기를 바란다. 하지만, 통계자료는 공산품과는 다소의 차이가 있다. 즉 통계자료의 공급자인 통계기관은 응답자 보호라는 외형으로 보이지 않는 조건이 있으므로 이를 감안하여 보급정책을 펴는 반면, 수요자인 심층 이용자는 통계자료도 공산품과 같이 고품질의 자료, 즉 원자료(original data) 수준의 마이크로 데이터를 쉽고 편리하게 활용할 수 있기를 바라는 것이 현실이다. 이와 같이 통계자료의 공급자와 수요자간의 관점의 차이를 정책적 또는 기술적으로 완화함으로써 공공재인 통계자료의 활용을 극대화할 수 있을 것이다. 그러므로 이들 간의 관점의 차이를 살펴본다.

1. 통계자료에 대한 공급자 관점

통계자료의 공급자인 통계기관의 가장 우선으로 취급하는 요소는 응답자와의 신뢰유지를 위하여 응답자의 개인정보 보호[10]에 있다. 이는 통계조사의 연속성을 유지하기 위하여 통계기관이 반드시 지켜야할 의무사항으로써 통계법에 응답자 보호 조항을 명시하여 통계작성에 참여한 모든 사람이 이법을 준수하도록 하고 있다. 이와 같이 응답자 보호가 통계기관에게는 중요 문제인 까닭에 통계자료 이용 활성화에는 다소 인색하고 소극적이었다.

그러나, 최근 들어 통계기관은 자신들이 작성한 통계자료를 공공재로 인식하고, 심층 이용자들의 적극적인 이용을 바라고는 있지만, 다음과 같은 부분에 믿음을 갖지 못하는 이유로 아직도 소극적이다.

첫째, 마이크로 데이터의 품질이 세분된 분석에 충분한지 여부, 특히 표본조사에서 충분하지 않는 표본 규모에 우려성이 높음. 또한 세분된 분석으로 높아지는 개별 정보유출 우려 등

둘째, 법적 또는 기타 수단으로 심층 이용자의 마이크로 데이터의 접근을 충분히 지원하고 통제할 수 있는 권한 유무문제

셋째, 비밀보호를 위하여 필요한 조치인 익명파일생성과 이의 접근을 위한 접속장치 및 보안장치 등의 도입에 소요되는 비용과 기술력 부족

넷째, 익명파일에 적용하는 노출통제(disclosure control)기술에 대한 신뢰도 등

이상과 같은 이유에서 통계기관은 마이크로 데이터의 보급에 다소간의 통제를 하고 있으며, 가능하면 비밀노출 위험이 우려되는 부분은 오히려 적극적으로 보급을 통제하려는 경향이 강하다. 이러한 경향은 한국은 물론 통계 선진국이라 할 수 있는 미국, 일본, 캐나다, 호주, 덴마크 등에서 유사하게 나타나고 있다. 그러나 이들 국가들은 [그림7]에서 보는 바와 같이 가급적 원 자료

10) 사업체통계조사의 경우는 기업의 비밀보호를 포함한다.

(original data)의 속성을 그대로 수요자에 전달하려는 연구를 오래전부터 진행하고 있고[1][2][3][4], 또한 일부는 통계작성표준지침[12]에 도입하는 등 적극적으로 움직이고 있다.

2. 통계자료에 대한 수요자의 관점

통계자료 수요자, 특히 심층이용자는 대체적으로 대학, 연구소, 정부기관 등에서 근무하는 학자들이 주류를 이루고 있으며, 최근에는 국제기구에 근무하는 직원도 증가하고 있는 추세이다. 이는 마이크로 데이터 이용자들이 얻을 수 있는 다음과 같은 장점이 있기 때문이다.

첫째, 마이크로 데이터를 이용하지 못하는 심층 분석자가 직접 자료를 수집할 경우 사회적 부담이 될 뿐만 아니라 수집한 자료가 표본축소로 인하여 품질이 떨어질 수 있다. 그러므로 심층 분석자들은 통계기관에서 서비스하는 마이크로 데이터에 대하여 많은 요구를 하는 것이다.

둘째, 국가 및 지방정부, 민간 등의 정책입안자가 복잡한 문제를 제기하거나 정책 추진에 깊은 분석이 필요할 경우 이미 생산되어 있는 마이크로 데이터를 활용할 수 있어 시간과 비용을 줄일 수 있는 장점이 있다.

셋째, 심층 이용자들이 마이크로 데이터 이용과 함께 피드백을 통하여 해당 데이터의 품질 개선이 가능해지는 장점도 있다.

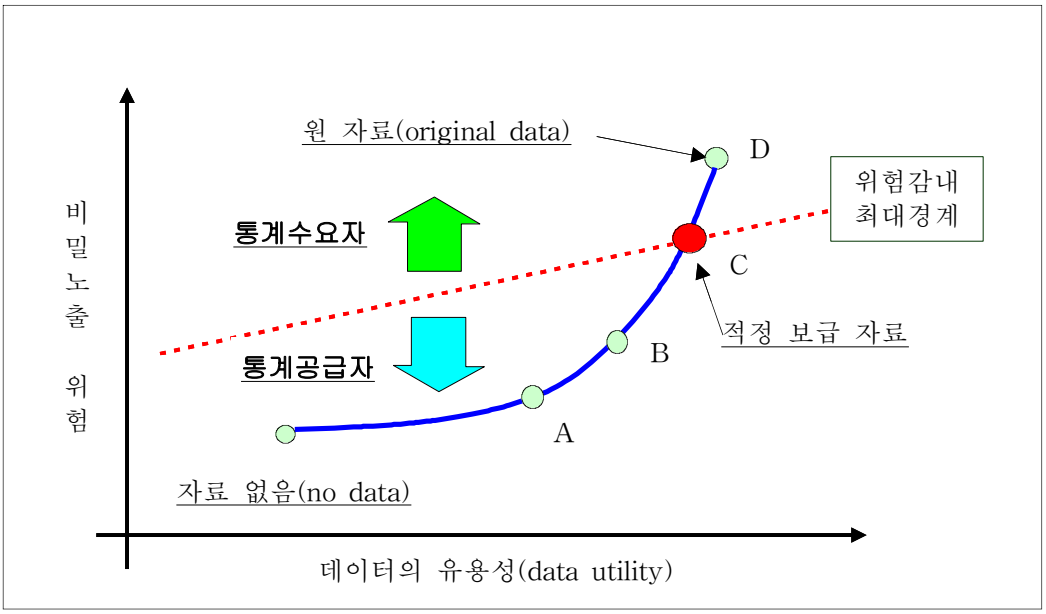
넷째, 통계 산출물을 증가시켜 전체 통계조사의 가치를 확대하는 효과가 있다. 이는 통계기관에서만 분석한다면 한계가 있지만, 여러 심층 분석자들이 여러 목적으로 분석하고 활용한다면 그의 사회적 가치는 엄청나게 증가하는 것은 당연한 것이다.

이상과 같이 마이크로 데이터의 이용을 확대함으로써 얻을 수 있는 장점이 많음에도 통계기관은 **응답자 보호의 목적, 데이터의 신뢰도 부족, 또는 조**

사목적 외에 이용하는 오용 문제 등을 우려하여 마이크로 데이터의 보급에 소극적이다. 이는 전 세계 거의 모든 통계기관에서 나타나는 현상으로 알려져 있다. 이러한 까닭에 심층 이용자들은 통계기관의 마이크로 데이터에 접근허용이 상당히 보수적이라고 느낌을 가지고 있다. 또한 공급자와 마찬가지로 데이터 활용을 증가 시킬 수 있는 방법에 대한 연구의 중요성을 수요자들도 인정하고 있다[6].

3. 통계자료에 대한 공급자와 수요자간의 긴장관계 해소 방법

앞에서 알아 본 바와 같이 통계자료에 대한 공급자와 수요자간 관점의 차이는 있지만, 상호간의 입장을 이해 못하는 것은 아니다. 공급자는 통계자료의 활용확대로 인한 장점을 모르지 않고, 수요자는 비밀보호 등과 같은 공급자의 어려움을 이해 못하는 것은 아니다. 그렇지만, 명쾌한 해결방안 또한 쉽지 않은 것이 현실이다. 해결방안 모색을 위하여 양자간의 입장을 보다 구체적으로 살펴보기 위하여 [그림7]을 활용한다.



[그림7] 통계 공급자와 수요자의 관점 관계

[그림7]은 통계자료 공급자와 수요자간 관점의 명확한 이해를 위하여 이들

간의 관계를 표현한 그림이다. X축은 데이터의 유용성을 나타내는 축이며, Y축은 비밀노출 위험성을 나타내는 축이다. 또한 A, B, C, D는 임의의 자료를 나타내는 지점으로 D는 원 자료, C는 최소한의 비밀보호 조치만 취한 마이크로 데이터, B는 별도의 주문에 의하여 작성된 통계표 수준의 자료이며, A는 완전 공개된 통계표 수준의 자료로 가정한다. 그리고 점선은 감내할 수 있는 비밀보호 위험 수위를 나타내는 선이다.

[그림7]에서 보는 바와 같이 가장 완벽한 유용성을 가진 통계자료는 D지점이며, 가장 가치 없는 자료의 지점은 물론 자료를 보급하지 않는 지점이다. 이러한 여러 지점 간의 관심 지점은 C지점인 "적정보급자료"가 된다. 공급자는 가능한 C지점을 B지점에 가깝게 이동하려는 경향이 강하고, 수요자는 C지점을 D지점으로 가깝게 하여 데이터의 이용가치를 극대화하려고 한다. 물론 A와 B 지점의 자료 또한 수요자는 가급적으로 A는 B 쪽으로, B는 C쪽으로 하려고 할 것이다. 그렇지만, 수요자의 관심은 A, B 자료보다는 C자료를 보다 D쪽으로 이동시키는 것에 더 관심이 많다.

[그림7]에서 나타난 바와 같이 공급자는 비밀노출 위험을 최소화하기 위하여 일반적으로 위험회피정책을 선호하게 되고, 이용자는 이러한 정책에 불편을 느끼게 됨으로써 통계기관의 통제를 불필요한 관료주의로 간주하는 등 상호간에 해결할 수 없는 긴장관계가 지속되고 있다. 이러한 긴장관계의 해소는 공급자가 사고의 전환 또는 정책의 전환 없이는 해결이 불가능하다. 따라서 **통계기관은 기존의 위험회피 방법에서 위험관리전략으로 전환이 필요하다.**

최근에 정보통신기술의 발전과 DB의 급속한 확장으로 인하여 심층이용자들은 인터넷, 각종 DB등을 통하여 수많은 자료를 손쉽게 확보할 수 있다. 반면, 상대적으로 통계자료의 경우는 이용자들이 통계자료에 대한 접근방법과 경로가 복잡하고, 그 내용도 요구수준 이하의 것만 확보할 수 있다고 불만을 토로하기도 한다. 이상과 같이 정보통신기술의 발전으로 정보의 보급 및 확보 경로가 단순화되고 간편해 지고 있는 사회적 변화에 통계자료 보급 또한 능동적으로 대처하여야 하는 것은 시

대적 흐름이라 할 수 있다. 따라서 기존의 통계자료 보급방법으로 선호하였던 위험회피 방법은 수요자의 욕구를 충족시키기 어렵기 때문에 보급자료 범위를 확대하는 대신 위험요소들을 분석하고 관리하는 전략으로 전환하여 통계자료의 활용을 확대하는 방향으로 정책전환이 필요하다. 다만, 이를 위하여 정책적, 기술적인 방법에 대한 연구와 준비가 선행되어야 한다.

제 3절 통계자료 이용 활성화를 위한 정책과 비밀보호

통계자료의 활용범위 확대는 비밀노출 위험성이 증가하는 문제와 오용 문제를 해결하지 않고서는 불가능한 일이다. 따라서 통계자료 활용확대 정책과 이로 인한 역기능, 즉 비밀보호 정책, 통계자료 오용 등에 대한 해결 방안을 동시에 강구하여야 한다. 이를 위한 방법은 다음과 같은 것을 들 수 있다.

첫째, 통계자료의 이용 시에 비밀보호를 위하여 법 제도적인 부분을 명확히 하여 이용자 실수 등으로 인한 비밀노출 사고에 대한 법적인 책임을 명확히 한다.

둘째, 통계자료의 오용을 방지하기 위하여 마이크로 데이터의 활용방법 설명자료도 함께 제공하여 이용자로 하여금 통계자료를 충분히 이해하고 활용하도록 유도하고 지원한다.

셋째, 원격접속제도(RAF)와 데이터 실험실(DL) 운영을 확대하여 이용자들의 마이크로 데이터 접근을 용이하게 하여야 한다. 이를 위하여 익명 마이크로 자료파일(AMF)의 생성기술을 연구 개발하여 높은 수준의 데이터 유용성을 유지하면서 비밀노출은 최소화한 자료를 생성하여 서비스하여야 한다.

이상의 3가지 사항 중 첫째와 둘째는 일반적인 사항이며, 원격접속제도와 데이터 실험실 운영 제도는 통계자료 이용 활성화에 핵심이 되는 정책적 제도이다. 따라서 향후에 연구발전이 기대되는 분야이다.

1. 원격접속제도(RAF)

원격접속제도(Remote Access Facilities)는 사전에 비밀노출을 방지하는 기술을 적용한 통계자료인 매크로 데이터 파일, 또는 마이크로 데이터 파일을 인터넷 등 통신망을 통하여 이용자가 원격지에서 접속하여 사용하는 방법이다. 이 방법은 최근에 정보통신기술의 발전과 인터넷 이용자가 증가하고 있어 통계자료 이용자들이 선호하는 방법이다.

이 제도는 국가별로 도입하는 방법이 다소 상이하며, 아직 초기 단계에 있다. 캐나다와 호주의 경우는 메일을 통하여 자료를 활용하는 전산 프로그램을 통계기관에 송부하여 처리결과를 메일로 돌려받는 **오프라인 방식**을 활용하고 있다. 반면 덴마크는 온 라인으로 이용자가 직접 처리하여 결과를 화면으로 출력하는 **온 라인 방식**을 활용하고 있다. 이와 같이 2가지의 마이크로 데이터의 원격접속방법이 있으며, 대다수의 국가들이 오프라인 방법을 취하고 있으며, 온 라인 방법은 아직 개발단계에 있다. 그러나 엄격한 의미에서 원격접속제도는 온 라인 방식을 말하며, 오프라인 방식은 주문통계제도와 원격접속제도를 병합한 방법이라 할 수 있다.

원격접속제도는 데이터 실험실(DL) 보다는 감독비용이 저렴한 장점이 있으나, 원격지에서 직접 원 자료(original data)에 접근하는 방법이므로 비밀노출 위험이 높은 방법이다. 따라서 다음과 같은 보안방안이 필요하다.

첫째, 비밀노출 방지를 위하여 자료에 비밀노출 방지기술을 적용하여야 한다.
둘째, 마이크로 데이터의 다운로드를 불가능하게 하여야 한다.
셋째, 데이터 활용규칙 위반시에 이용 제재 등 위험관리를 강화하여야 한다.
넷째, 실명화 위험(identification risk)은 항시 존재하므로 정기적인 감독과 이용자의 자료이용 규칙을 숙지시키기 위한 조치를 취하여야 한다.

이상과 같은 보안방안은 온라인 방식에는 필수적인 사항이며, 이러한 준비 없이 원격접속제도를 시행할 수 없는 것은 너무나 당연하다.

2. 데이터 실험실(DL)

데이터 실험실 제도는 통계기관에서 오래 전부터 마이크로 데이터를 대외 이용자에게 서비스하는 제도로써 활용하여 왔다. 이 제도는 통계자료 이용자가 해당 데이터가 있는 통계기관의 데이터 실험실에 직접 방문하여 통계기관의 감독 하에 필요한 통계자료를 자유롭게 사용하도록 하는 제도이다. 이 제도의 운영을 위해서는 다음과 같은 요건이 구비되어야 한다.

첫째, 통계자료의 데이터 실험실 운영내용과 활용 방법이 홍보되어야 한다.
둘째, 통계기관 전담직원의 철저한 감독이 필요하다.
셋째, 외부 통신망과의 접속을 철저히 통제하여야 한다.
넷째, 출력물의 외부반출은 정당한 자료만 가능하며, 기타 자료의 반출은 철저히 통제하여야 한다.
다섯째, 이용규정을 마련하여 위반 시에 제재를 명확히 하고, 강화하여야 한다.

이상과 같은 수준의 보안대책이 필요하다. 데이터 실험실 제도는 이용자가 자신의 연구목적에 접합하도록 자유자재로 자료를 취급할 수 있다는 장점은 있지만, 이용자가 익숙하지 않은 소프트웨어를 사용해야 하는 전산작업 환경이 불편한 문제가 있다. 또한 다른 제도에 비하여 관리비용이 높다는 것이 통계기관의 고민이 될 수 있다. 그러나 마이크로 데이터의 다양한 분석을 통계작성기관에서 지원하기 어려운 점, 또한 데이터의 유용성이 높은 자료, 즉 가치있는 자료의 외부 반출이 불가능 점 등으로 인하여 수요자와 공급자가 선호하는 제도라 할 수 있다.

제 4 장 통계자료의 비밀보호를 위한 보안기술

제 1절 보안침해 유형과 일반적 보안대책

넓은 의미의 통계자료 보안은 미래에 예상되는 바람직하지 못한 사건 또는 활용으로부터 효율적으로 대비하는 것을 말한다. 이를 위한 방법으로는 물리적 대책, 관리적 대책, 기술적 대책, 법·제도적 대책이 있으며, 이들 방법을 상호 연계한 종합적인 대책이 바람직한 통계자료 보호를 위한 방법이다.

최근 들어 인터넷 기반의 컴퓨터 사용자가 증가함에 따라 개인정보누출에 따른 사생활 침해문제가 사회적으로 문제가 되고 있고, 컴퓨터 바이러스에 의한 전산망의 성능저하, 자료 파괴 및 수정 등으로 인한 재산상의 손실문제, 또한 개방형 시스템을 기반으로 하는 인터넷의 활용 활성화는 해킹 및 스팸메일 등과 같은 다양한 정보화 역기능이 있다. 이와 같이 정보화 진전에 따라 나타나는 역기능이 통계자료에서의 비밀보호 문제와는 무관하지 않다. 그 이유는 통계자료에 개인 비밀사항이 포함되어 있고, 이 자료들이 인터넷 등 통신망을 통하여 보급 및 활용되고 있기 때문이다.

따라서 통계자료에서의 비밀보호 문제를 넓은 의미의 보안대책에서부터 접근하여 통계자료 영역으로 확장하는 방법으로 통계자료의 비밀보호 방법을 고찰한다.

1. 보안침해의 종류 및 유형

보안침해의 종류에는 다음과 같이 3가지 유형이 있다.

첫째, 물리적인 보안침해 유형으로 물, 불, 전기 등에 의해 보관중인 통계자료가 파손 또는 망실되는 경우를 말하며, 이의 대비책은 통계자료의 소산관리 및 재난복구시스템(DRS : Disaster Recovery System) 구축 등이 있다.

둘째, 소프트웨어에 의한 보안침해가 있으며, 이는 컴퓨터 바이러스에 의하여 보관중인 통계자료의 수정, 삭제 또는 변경 등이 인하여 자료가 손상되는 침

해 유형이 있다. 이에 대한 대책은 최신 컴퓨터 바이러스 백신 프로그램 활용으로 가능하지만, 이 방법이 간단한 것 같지만 실은 귀찮고 불편하므로 소홀히 하는 데서 문제를 일으키고 있다.

셋째, 데이터에 의한 보안침해 방법으로 출력중인 데이터를 모니터링 하여 활용하는 유형이 있다. 이 침해에 대한 대책은 비밀자료의 출력실에 외부인 출입을 제한하는 방법이 있다.

넷째, 익명화된 통계자료를 기타자료와 연계하여 실명으로 전환하는 데이터 링케이지(data linkage) 방법으로 개인의 비밀이 노출되는 경우가 있다. 이 경우 원래 목적은 통계자료의 가치를 높이는 방법이나 작업과정에서 부수적으로 발생하는 실명화로 개인의 비밀이 노출될 우려성이 있다. 이러한 경우에는 마이크로데이터의 매트릭스 마스킹 방법으로 해결하고 있다[2].

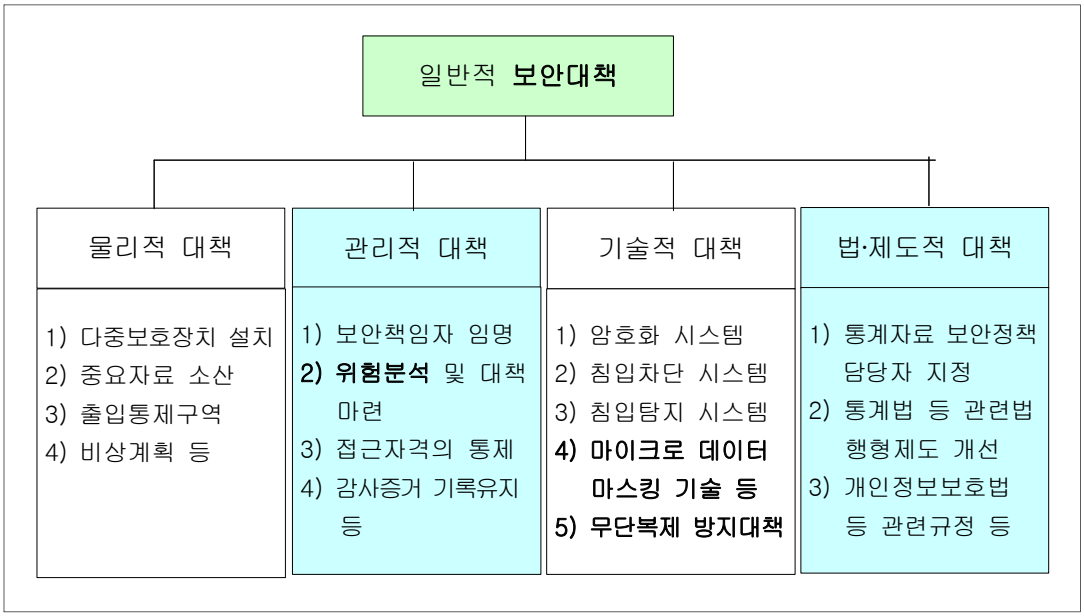
이상과 같이 4가지 유형 모두가 통계자료의 보호를 위한 대책이 필요한 침해유형이다. 이들 중에서 통계자료의 이용 활성화를 위해 필요한 침해유형은 네 번째의 마이크로 데이터 링 케이지로 인한 실명화 문제이다.

2. 일반적 보안대책

앞에서 언급한 바와 같이 통계자료의 근본적인 보안대책은 크게 물리적, 관리적, 기술적, 법·제도적 보안대책으로 구분할 수 있다.

[그림8]에는 이러한 대책들에 대한 세부적인 사항을 나타내고 있다. **물리적 대책**은 실제 통계자료의 물리적인 부분에 대한 보호정책이며, 여기에는 전산장비의 이중화, 통계자료 테이프의 이중 백업 및 소산, 출력실 및 전산실의 출입통제 등의 방법이 있다.

관리적 대책에는 보안책임자의 임명, 위험분석 및 대책마련, 접근자격의 통제 등과 같이 관리적인 측면의 보안대책을 말한다.



[그림8] 통계자료에 대한 일반적 보안대책

기술적 대책은 통신망의 해킹을 방어하는 침입차단시스템, 침입탐지시스템, 송수신되는 자료의 외부 노출시에도 보호하기 위한 암호화 시스템 등이 있다. 또한 개인의 비밀노출을 방지하기 위한 실명화 방지기술로 마이크로 데이터 마스킹 기술, 무단복제 방지를 위한 워터마킹 기술 등이 있다.

법·제도적 대책에는 통계자료의 보안정책 담당자를 지정하여 개인의 비밀노출은 물론 마이크로 데이터의 대외 서비스 정책을 개발하고 추진하는 업무를 담당하는 제도, 통계법에서의 통계자료의 행형제도 개선 등을 예로 들 수 있다.

이상과 같이 일반적 보안대책은 정보보호 또는 비밀보호를 위한 종합적인 대책이며, 마이크로 데이터의 활용 활성화를 위하여 고려해야 하는 비밀보호 대책은 데이터의 접근 방법에 따라 차이가 있다. 예를 들면, 원격접속제도의 경우에는 해킹방어를 위한 대책, 비밀노출 방지 대책이 필수적이며, 추가적으로 법 제도적인 대책도 요구된다. 다음 절에서 마이크로 데이터 서비스 제도별로 보안대책 유형에 대하여 알아본다.

제 2절 데이터 유형별 보안위험 분석

마이크로 데이터의 대외 서비스 유형은 <표4>에서 보는 바와 같이 5가지 유형이 있다. 이들 유형별로 실명화 방지 측면에서 고려해야하는 보안 대책을 알아본다. 이를 위하여 우선 <표4>의 유형마다 취급하는 마이크로 데이터의 형태를 먼저 살펴보고 보안 위험요인을 분석하면 <표5>와 같다.

<표5> 마이크로 데이터 서비스 유형별 보안위험

서비스 형태		제공되는 자료 내용 또는 형식	보안위험
통계표(tables) 형태로 가공하여 대외 서비스하는 방법 (주문 통계서비스 제도)		◦ 통계표 : 매크로 데이터 수준(A형)[11]	
		◦ 통계표 : 마이크로 수준으로 세분된 통계표(B형)	있음 (비밀노출)
익명 마이크로 자료파일 (AMF)	공공이용 파일 (public use file)	◦ 마이크로 데이터(C형) ◦ 비밀보호 기술을 적용하여 실명화 위험을 최소화하는 것이 필요한 자료	있음 (무단복제)
	인가파일 (licensed file)	◦ 특수 목적용 마이크로 데이터 ◦ 목적에 따라 두 가지 유형의 자료파일 제공 -실명화를 방지한 파일 형태자료(D형) -원 자료(original data)(E형) ◦ 원 자료의 경우는 이용자의 특별인가가 필요하며, 법적인 조치가 필요	있음 (법적문제)
원격접속제도(RAF)		◦ 컴퓨터 통신망을 통하여 이용자들이 마이크로 데이터를 이용할 수 있도록 하는 제도 ◦ 서비스 가능한 자료 : B형, C형, D형	있음 (비밀노출) (해킹우려)
데이터 실험실(DL)		◦ 모든 유형의 자료 ◦ 철저한 관리 및 감독이 필요한 제도	있음 (법적문제)

<표5>에서 보는 바와 같이 서비스 형태별 마이크로 데이터 유형은 A형, B형, C형, D형, E형 등 5종으로 구분되며, 이들 자료의 서비스 방법에 따른 특별한 방법으로 원격접속제도와 데이터 실험실 운영제도가 있다.

---

11) 서비스 자료의 구분을 위하여 A형, B형, C형, D형, E형으로 표기

1. 매크로 데이터의 보안위험 분석

매크로 데이터는 일반적으로 개인 비밀노출의 위험, 즉 실명화 위험은 거의 없다고 하지만, 특수한 경우에는 실명화 우려가 있는 경우가 있다. 통계조사의 경우 실명화 방지는 응답자 보호를 의미한다. 매크로 자료 통계표(table)의 세분정도에 따라 응답자 정보의 실명화가 가능한 경우가 있다. 빈도가 1, 2, 3인 경우에 통계표에서 직접 또는 간접으로 개인정보의 실명화로 비밀이 노출될 수 있다.

예를 들어 특수한 산업분류의 사업체 수가 2개인 경우에 a_1 사업체와 a_2 사업체가 상호 경쟁업체라면 상대방의 매출 등 정보가 중요한 정보로 활용될 수 있다. 이 경우 a_1 업체가 자신의 정보를 제외한 자료가 a_2 업체에 대한 정확한 정보를 쉽게 확보할 수 있다.

<매크로 자료)

사업체 수	종업원 수	매출액(억원)	수출액
…	…	…	
2	150	200	
…	…	…	

< a_1 자신의 자료>

1	80	95	

< a2에 대한 노출 정보 >

1	**70**	**105**	

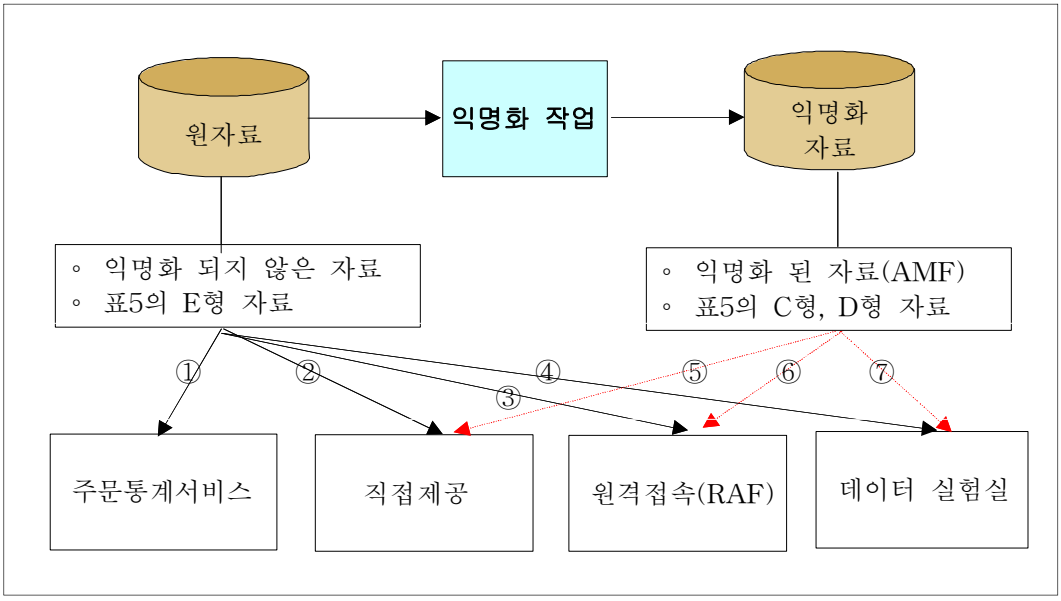
이와 같이 매크로 데이터의 경우라도 빈도가 2인 경우에는 실명화가 가능한 문제가 있다. 동일한 방법으로 빈도 3인 경우도 2개 자료만 알 수 있다면 나머지 1개 응답자에 대한 정보의 실명화가 가능하며, 마찬가지로 4개인 경우에도 그러한 방법으로 실명화가 가능하다.

이와 같이 빈도에 따른 실명화를 위해서는 빈도가 2인 경우 1개 응답자의 추

가정보, 3인 경우는 2개의 추가정보, 4인 경우는 3개의 추가정보, 5인 경우는 4개의 추가정보가 필요하다. 이와 같이 추가정보의 요구정도에 따라 비밀노출 위험 정도가 달라진다. 1인 경우가 가장 위험하고, 다음 2, 3, 4, 5로 위험의 정도가 약해진다. 그러므로 매크로 데이터는 세분화 되면 될수록 위험도가 높아진다고 볼 수 있다. 따라서 자료를 세분할 때, 범주속성의 빈도를 살펴서 실명화 위험의 정도가 감내할 수 없는 수준(일반적으로 빈도 3이하 또는 5이하로 알려져 있음)에 대하여는 실명화 방지방법이 필요하다.

2. 마이크로 데이터의 보안위험 분석

마이크로 데이터는 매크로 데이터에 비해 비밀노출 위험이 상대적으로 높다. 마이크로 데이터는 익명화 되지 않은 원 자료(original data)와 익명화된 자료로 구분되며, 이들은 서비스 방식 또는 제도에 따라 실명화 위험성은 다르게 나타난다. 따라서 보안위험을 알아보기 위하여 [그림9]에 나타난 바와 같이 자료유형과 서비스 제도와의 관계를 먼저 살펴본다.



[그림9] 자료 유형과 서비스 제도의 관계

[그림9]는 원 자료(original data)와 이를 익명화한 AMF형태가 있으며, 이들 자료와 서비스 형식 4가지의 관계를 나타내는 그림이다. 여기서 원 자료와 익명화 자료에 대하여 7가지의 서비스 유형, 즉 ①부터 ⑦까지의 종류가 있다. 그러므로 자료 자체의 위험성과 서비스 방법상의 위험성을 구분하여 위험요인을 분석한다.

(1) 익명화 되지 않은 원 자료의 위험분석

익명화 되지 않은 원 자료는 실명 자료이기 때문에 철저한 관리와 통제가 필요한 자료이다. 그러므로 서비스하는 방법 4가지에 대하여 실명화 위험정도를 분석한 결과 서비스 가능여부는 <표6>과 같다.

<표6> 원 자료의 보안위험 분석표

서비스 방법	보안위험분석	서비스가능여부	보안대책
①주문통계서비스	◦ 위험도 : 매크로 데이터와 동일 ◦ 빈도에 따른 실명화 위험존재	◦ 가능	◦ 빈도에 따른 노출방지대책
②직접제공(CD 등)	◦ 위험도가 상당히 높은 방법 ◦ 특수목적 서비스는 인가제도 필요	◦ 원칙적으로 불가	◦ 인가제도 ◦ 복사방지
③원격접속(RAF)	◦ 위험도 상당히 높은 방법 ◦ 해킹 등 위험도가 매우 높음	◦ 불가	-
④데이터실험실	◦ 위험도는 높지만, 철저한 감독하에서 위험도 감소 가능	◦ 가능	◦ 관리적 대책

위 표에서 보는 바와 같이 ①의 방법은 원 자료를 이용자의 요구사양에 따라 통계기관에서 통계표를 작성하여 그 결과를 제공하는 방법이므로 매크로 데이터를 제공하는 것이다. 그러므로 실명화 위험도는 거의 없지만, 앞에서 논의한 빈도에 따른 비밀노출 위험은 존재하므로 이에 대한 대책만으로 해결이 가능하다.

②의 방법은 마이크로 데이터를 CD-ROM 등 매체에 수록하여 서비스하는

방법이므로 불특정 일반인에 대한 서비스는 위험하므로 제한하여야 한다. 그러나 통계조사를 위한 표본추출용 모집단 자료와 같이 특수한 경우에는 부득이 서비스를 해야만 하는 경우가 있다. 이 때는 사용자 인가관리와 함께 제공되는 마이크로 데이터의 무단복제 사용을 방지하는 것이 필요하다. 복제방지는 제공된 자료의 불법적인 유통으로 발생할 수 있는 비밀노출 문제 발생시에 나타나는 책임소재 문제와 서비스 목적외의 사용으로 발생하는 오용문제 등을 야기할 수 있고, 또한 유료 보급된 자료의 불법복제 사용은 통계제공 관련 정책으로는 통제할 수 없다. 따라서 워터마킹(water marking) 기술 또는 이의 응용한 기술을 활용하여 불법복제를 방지하여야 한다.

③의 방법은 익명화 되지 않는 자료를 온라인으로 원격지에서 접속하여 이용자가 직접 활용하는 방법으로 상당히 위험하므로 서비스는 불가능하다, 다시 말하여 불특정 다수에 실명 자료를 온라인상으로 공개하는 것은 현행 법 테두리 내에서 허용되지 않는 방법이다. 그러나 원격접속제도의 개념에 대하여 약간의 논란이 있으므로 정리가 필요하다. 전자 메일 등을 통하여 관련 프로그램과 주문서를 통계기관에 송부한 후 처리결과 통계표를 돌려받는 방법을 오프라인 방법의 원격접속제도 범주에 포함하는 경우가 있다. 본 논문의 3장 3절에서 캐나다와 호주의 예를 소개한 바 있다. 자세히 보면, 이 방법은 원격접속제도보다는 주문통계서비스와 데이터실험실 개념을 혼합한 방법으로 볼 수 있다. 그러므로 이 방법의 위험도는 주문통계서비스 제도와 동일하다.

④의 방법은 이용자가 직접 통계기관의 특정장소에 방문하여 필요한 마이크로 데이터를 신청하여 작업한 후, 결과만을 외부에 반출하는 제도이다. 이 제도는 마이크로 데이터가가 필요하지만 비밀노출 문제가 있어 외부로 반출할 수 없는 문제를 해결하기 위한 대안으로 사용하는 것이다. 그러므로 보안위험은 존재하지만, 외부에서 방문한 작업자를 철저한 감독으로 해결이 가능한 것이 특징이다, 감독 방법은 감독자를 임명하여야 하고, 또한 작업에 사용하는 장비는 메일 등 외부로 자료를 송부할 수 있는 경로를 차단하는 것이 필수적이다.

(2) 익명화 자료의 위험분석

익명화 자료는 원 자료(original data)를 개별 응답자에 대한 정보식별이 불가능하도록 변화시켜 만들어진 대외 서비스용 자료를 말한다. 익명화 자료는 익명화 작업 방법에 따라 데이터 링케이지 등을 통한 실명화 위험이 없지는 않지만, 이러한 문제는 다음 절에서 취급하기로 하고, 여기서는 익명화 자료를 이용자에게 서비스하는 방법에 따른 위험성을 분석한다.

[그림9]를 보면, 익명화 자료를 이용자가 접하는 방법은 4가지가 있지만, 주문통계서비스 방법은 원 자료의 서비스와 동일 수준 또는 그 이하의 실명화 위험이 있으므로 생략하고, 잔여 3가지 방법, 즉 ⑤, ⑥, ⑦의 방법에 대하여 분석한 결과가 <표7>과 같다.

<표7> 익명화 자료의 보안위험 분석표

서비스 방법	보안위험분석	서비스가능여부	보안대책
⑤직접제공(CD 등)	◦ 위험도 : 익명화 기술에 따름 ◦ 무단복제 위험성 존재	◦ 가능	◦ 무단복제방지
⑥원격접속(RAF)	◦ 다운로드 위험성 존재 ◦ 해킹 위험성 존재	◦ 가능	◦ 다운로드방지 ◦ 해킹방지
⑦데이터실험실	◦ 작업 자료의 외부 직송 위험 ◦ 자료 복제 반출 위험	◦ 가능	◦ 철저한 감독 ◦ 외부통신단절

⑤의 방법은 익명화 자료를 CD-ROM 등과 같은 매체에 수록하여 서비스하는 방법으로 실명화 위험성은 익명화 작업 방법에 따라 다소간의 차이는 있지만, 거의 없다고 볼 수 있다. 그러나 외부에 제공된 CD-ROM을 무단으로 복제하여 신청목적 외 다른 용도로 사용할 위험성은 항시 존재한다. 그러므로 무단복제 방지를 위한 기술의 적용이 필요하다.

⑥의 방법은 원격지에서 이용자가 직접 익명화 마이크로 데이터를 접속하

여 처리·활용하는 방법이다. 이 방법의 위험성은 이용자 또는 임의의 접속자가 올바르지 못한 의도로가 자료를 다운로드 받거나 해킹할 우려성이 있다. 따라서 이러한 위험성을 방지하는 대책이 요구된다.

⑦의 방법은 이용자가 통계기관의 지정장소에 방문하여 익명화 마이크로 데이터를 활용하는 방법으로 특별한 보안 위험성은 없지만, 작업자의 의도에 따라 마이크로 데이터를 이메일 등 통신망을 통하여 직접 외부로 반출하는 위험성과 복사하여 반출할 위험성이 있다. 따라서 외부와의 통신선로를 차단하여 외부로 반송을 방지하여야 하고, 철저한 감독을 통하여 복사 반출을 방지하여야 한다.

3. 익명화 자료 자체의 위험분석

익명화 자료는 원 자료를 익명화 작업을 거쳐 이용자에게 서비스용으로 만들어진 마이크로 데이터를 말한다. 익명화 작업은 여러 가지 방법이 있으며, 최적의 방법은 비밀노출 위험성을 최소화하고, 자료의 유용성(utility)을 극대화하는 방법이 된다.

익명화 자료와 다른 자료를 연계하여 익명화 자료를 실명으로 전환하는 데이터 링케이지(data linkage) 방법이 알려져 있다[2][3]. 이 방법은 익명화 되었다고 무심코 외부 이용자에 제공할 경우 <표8>과 같은 허점이 있어 실명으로 전환되는 위험성이 있다.

<표8> 데이터 링 케이지 예

identifier	Key	Key
이름	나이	주소코드
AA	63	32
BB	51	23
CC	37	11
DD	19	21
EE	25	12

Key	Key	Protected Variable
나이	주소코드	수입(천원)
37	11	89
43	11	46
43	11	52
37	12	55
31	12	40

<표8>을 보면, 왼쪽 표와 오른쪽 표를 각각 이용자가 확보하였다면, 키(key) 부분 각각 2개(총 4개)를 이용하여 다음과 같은 실명 자료를 확보할 수 있다.

"CC라는 사람의 나이가 37세, 11지역에 살고 있으며 수입 **89천원**이다"

<표8>의 내용을 볼 때, 보호되어야 하는 것 정보는 **"CC = 89천원"**이다. 그러므로 실제 정보를 제공할 때는 <표8>의 오른쪽과 같이 익명으로 제공하였지만, 쉽게 확보할 수 있는 왼쪽 자료와 연계함으로써 실명으로 전환이 가능한 허점이 있다. 이외에도 여러 가지 허점이 있지만 대표적인 실명화 허점으로 데이터 링케이지가 알려져 있다. 이러한 허점을 익명화 작업 시에 고려하여야 한다.

4. 보안위험 분석결과 종합

앞에서 매크로 데이터 및 마이크로 데이터를 이용자에 서비스하는 방법별로 보안위험을 분석하였다. 그 내용을 다시 한번 종합하여 정리하면 다음과 같이 5가지로 압축할 수 있다.

<표9> 마이크로 데이터 서비스를 위한 비밀노출 방지대책 종합

비밀노출 방지대책	내 용	분야
◦ 빈도에 따른 노출방지	◦ 빈도 3이하에 대한 처리방안	◦ 매크로 데이터 서비스 분야
◦ 마이크로 데이터의 익명화 방법	◦ 실명화 불가능한 기술적 방법	◦ 마이크로 데이터 서비스 분야
◦ 무단복제 방지	◦ 자료수록 CD 등의 복제방지 기술	◦ 유료판매자료
◦ 다운로드 방지	◦ 온라인으로 이용자 접속한 자료의 다운로드 방지	◦ 원격접속제도 중 온라인 방법
◦ 데이터실험실 운영지침	◦ 방문 이용자 관리 등의 운영지침	◦ 데이터 실험실

이상과 같이 5개의 대책을 3절에서 자세히 알아본다. 이중에서 특히 익명화 기술에 중점을 두어 살펴본다.

제 3절 비밀노출 방지를 위한 보안기술

1. 빈도에 따른 노출 방지 방법

통계표의 민감 항목의 셀(cell)에 특정 응답자 수가 소수인 경우에는 비밀노출 위험, 즉 실명화 위험이 있다. 그 수는 통계자료에 따라 3 또는 5이하로 알려져 있지만, 일반적으로 3이하가 대부분이다. 이러한 위험은 다음과 같은 2가지 방법으로 그 위험을 제어하고 있다.

첫째 방법으로 범주(categories)를 조정하여 위험한 빈도가 나타나지 않게 하는 방법이 있지만, 이 방법은 일반적으로 범주속성의 특성인 계층구조에서 상위계층으로 이동시키는 방법이 일반적이다, 그 결과 통계표의 가치가 낮아지는 단점이 있다.

둘째 방법은 셀 값 감추기 방법(cell suppression), 랜덤 올림(random rounding) 방법, 제어 올림(controlled rounding) 방법, 비밀 편집(confidentiality edit) 방법 등을 사용하여 민감한 항목의 셀에 대한 실명화를 방지하고 있다[1].

아래 <표10>은 실명화 위험성이 있는 가상의 표이며, 이 표에는 비행청소년 관련 정보를 담고 있다. 응답자 빈도 5이하를 민감한 셀로 정의하고, 해당 셀에는 특수표기(*)를 하였다.

<표10> 교육정도 통계표(가상의 통계표 : 비밀노출 위험성 높음)

county	교 육 정 도				
	합계	낮음	중간	높음	아주 높음
합계	135	50	35	30	20
Alpha	20	15	1*	3*	1*
Beta	55	20	10	10	15
Gamma	25	3*	10	10	2*
Delta	35	12	14	7	2*

1.1 셀 값 감추기 방법(suppression)

민감 항목의 셀 값에 대한 노출방지를 위한 대표적인 방법으로 셀 값 감추기 방법(suppression)이 있다. 이 방법은 민감한 셀의 행과 열의 주변 값도 동시에 감추어야 한다. 여기에서 부수적으로 감추어지는 셀을 **보조 셀 감추기**(complementary suppression)이라 하며, 이 셀은 인위적으로 선정하여 처리한다.

아래 <표11>은 <표10>의 3이하 빈도 셀의 행과 열에 대하여 이웃하는 셀을 추가하여 감추기를 한 통계표이다. 이 통계표에 있는 민감한 셀 값의 비밀노출 방지가 되었는지를 알아보자.

<표11> 교육정도 통계표(감추기 셀 선정) [12] : **노출위험 존재**

county	교 육 정 도				
	합계	낮음	중간	높음	아주 높음
합계	135	50	35	30	20
Alpha	20	15	D_1	D_2	D_3
Beta	55	20	D_4	D_5	15
Gamma	25	D_6	10	10	D_7
Delta	35	D_8	14	7	D_9

<표11>은 실명화 위험은 제거되지 않았다. 행과 열의 선형조합으로 처리하면, 다음과 같은 결과를 얻으므로 아직도 민감 항목에 대한 비밀노출 위험은 존재한다.

$$(15+D_1+D_2+D_3) + (20+D_4+D_5+15)-(D_1+D_4+10+14)-(D_2+D_5+10+7)$$
$$= 20 + 55 - 35 - 30$$

위 식을 계산하면 $D_3=1$ 임이 노출된다. 이와 같이 보조 셀 감추기는 적용이 그리 쉽지 않은 방법이다. 그러므로 셀 감추기 방법은 보조 셀에 대한 선택이 중요한 요소로 작용한다. <표12>는 <표11>과 동일하게 총 16개 셀 중에서 7

12) D는 비밀노출 방지 셀을 나타낸 기호

개 셀의 자료는 공개하고, 잔여 9개 셀을 감추기 하였다. 그렇지만 그 결과는 <표11>은 비밀노출 위험이 있고, <표12>는 비밀노출 위험이 없는 것으로써 보조 셀의 역할이 중요하다는 것을 나타내고 있다.

<표12>교육정도 통계표(비밀노출 셀에 대하여 감추기 셀 조정) : **노출방지**

county	교 육 정 도				
	합계	낮음	중간	높음	아주 높음
합계	135	50	35	30	20
Alpha	20	15	D	D	D
Beta	55	20	10	10	15
Gamma	25	D	D	10	D
Delta	35	D	14	D	D

<표11>과 <표12>는 모두 16개의 셀 중에 9개 셀의 값을 감추어서 비밀노출 방지를 하는 방법이다. 그 결과 44%(9/16)의 자료가 감추어져 자료 자체의 가치가 떨어지는 문제점을 가지고 있다.

1.2 랜덤 올림 방법(random rounding)

미국 센서스국에서는 셀 값 감추기(suppression) 방법의 단점인 감추어지는 데이터 셀의 수를 줄이기 위하여 다른 방법을 연구하였는데, 그 대표적인 방법이 랜덤 올림(random rounding) 방법이다[1].

랜덤 올림 방법은 임의의 수를 기준으로 올림(round up) 또는 절삭(round down)하는 방법이다. 예를 들어 임의 수가 5라면, 각 셀의 값 X는 다음과 같은 형식으로 표현할 수 있다.

$$X = 5q + r$$

여기서 q는 양의 정수, r은 나머지이다. 이 경우에 가능한 나머지 값은 0, 1,

2, 3, 4가 된다. 올림의 값은 5*(q+1)가 되며, 그의 확률은 r/5이다. 절삭의 경우는 5*q가 되며, 그의 확률은 (1-r/5)이다. 이와 같은 방법으로 <표13>의 결과를 얻었다.

<표13> 교육정도 통계표(랜덤 올림 방법적용) : **노출방지**

county	교 육 정 도				
	합계	낮음	중간	높음	아주 높음
합계	135	50	35	30	20
Alpha	20	15	0	0	0
Beta	55	20	10	10	15
Gamma	25	5	10	10	0
Delta	35	15	15	10	0

<표13>에서 보는 바와 같이 랜덤 올림 방법을 적용한 결과, 실명화 위험성은 방지되었다. 그러나 행과 열의 합계가 맞지 않는 결과를 가져왔다. 즉 범주 값이 “낮음”의 경우 실제 합계는 55이지만, 결과는 50이다. 그리고 “아주 낮음”의 경우, 실제 합계는 15이지만, 결과 합계는 20으로 되어 있어 서로 맞지 않다. "Alpha"나 "delta"도 동일하다. 이와 같이 올림 방법은 행과 열의 합이 일치하지 않는 것이 단점이다.

1.3 제어 올림 방법(controlled rounding)

제어 올림 방법은 랜덤 올림 방법에서 행과 열의 합이 일치하지 않는 단점을 해결한 방법이다. 이 방법은 <표14>에서 보는 바와 같이 행과 열이 맞지 않는 것은 제어하여 일치시키는 방법이다. 이 방법은 행과 열의 합을 일치시키는 장점은 있지만, 몇 가지 단점을 가지고 있다. 첫째 단점은 컴퓨터 프로그램으로 구현하기 어렵다는 것이다. 그러므로 인위적으로 처리할 수밖에 없다. 둘째 단점은 복잡한 통계표에는 적용하기 어렵고, 또한 해결할 수 있는 방법

이 존재하지 않을 수 있다는 것이다. 이와 같은 이유로 아직 현장에서 사용하지 않는 방법이다.

<표14> 교육정도 통계표(올림 방법 적용) : **노출방지**

county	교 육 정 도				
	합계	낮음	중간	높음	아주 높음
합계	135	50	35	30	20
Alpha	20	15	0	5	0
Beta	55	20	10	10	15
Gamma	25	5	10	10	0
Delta	35	10	15	5	5

1.4 비밀 편집(confidentiality edit) 방법

비밀 편집방법은 1990년에 미국 센서스 국에서 개발한 새로운 방법으로 통계표(table)를 작성하기 전에 원 자료를 수정하여 매크로 데이터에서 발생하는 실명화 문제를 제어하는 방법이다. 이 방법을 활용할 때 유의사항은 조정된 원 자료(adjusted files)는 제공(release)하지 않으며, 오직 통계표 작성을 위하여 사용한다는 것이다.

비밀 편집 방법은 두 가지 유형이 있는데, 첫째로는 100% 마이크로 데이터 파일에 대하여는 데이터 교환(data swapping) 방법을 많이 사용한다. 둘째는 표본자료의 경우는 표본자료 자체가 실명화 위험성을 방지하고 있다고 본다. 다만, 좁은 지역의 경우는 추가적인 보호방법이 필요한데, 이는 지역적으로 좁기 때문에 표본자료가 없는 공백이 발생할 수 있기 때문이다. 이 경우는 대체 값(imputed value)을 활용한다[1].

비밀 편집 방법의 예를 100% 마이크로 데이터를 이용하여 보이고자 <표10>과 <표14>에서 사용한 데이터 중에서 <표15>와 같이 "Alpha" 카운티의

마이크로 데이터 파일은 활용한다.

<표15> 가상의 마이크로 데이터 : Alpha 카운티의 모든 레코드

번호	어린이	county	교육정도*)	소득	인종
1	John	Alpha	아주 높음	201	B
2	Jim	Alpha	높음	103	W
3	Sue	Alpha	높음	77	B
4	Pete	Alpha	높음	61	W
5	Ramesh	Alpha	중간	72	W
6	Dante	Alpha	낮음	103	W
7	Virgil	Alpha	낮음	91	B
8	Wanda	Alpha	낮음	84	W
9	Stan	Alpha	낮음	75	W
10	Irmi	Alpha	낮음	62	B
11	Renee	Alpha	낮음	58	W
12	Virginia	Alpha	낮음	56	B
13	Mary	Alpha	낮음	54	B
14	Kim	Alpha	낮음	52	W
15	Tom	Alpha	낮음	55	B
16	Ken	Alpha	낮음	48	W
17	Mike	Alpha	낮음	48	W
18	Joe	Alpha	낮음	41	B
19	Jeff	Alpha	낮음	44	B
20	Nancy	Alpha	낮음	37	W

*) 교육정도는 가구주의 교육정도를 말함

위의 <표15>를 이용하여 비밀 편집 방법의 예는 아래의 순서에 따른다.

① 4번과 17번 레코드가 선택되었다고 가정한다.
② 원래 카운티별 교육정도 통계표를 원하므로 임의의 다른 카운티에서 변수 인종, 성별, 소득을 연계하고, 추가적으로 10%(2개 레코드) 연계한다.
③ 연계 결과에 따라 원 자료를 편집하여 수정한다. 그 결과가 <표16>이다.
④ <표16>을 이용하여 작성한 통계표가 <표17>이다.

<표16> 가상의 마이크로 데이터 : 일부를 교환 편집한 결과[13)]

번호	어린이	county	교육정도	소득	인종
1	John	Alpha	아주 높음	201	B
2	Jim	Alpha	높음	103	W
3	Sue	Alpha	높음	77	B
4*	**Alfonso**	**Alpha**	**아주높음**	**61**	**W**
5	Ramesh	Alpha	중간	72	W
6	Dante	Alpha	낮음	103	W
7	Virgil	Alpha	낮음	91	B
8	Wanda	Alpha	낮음	84	W
9	Stan	Alpha	낮음	75	W
10	Irmi	Alpha	낮음	62	B
11	Renee	Alpha	낮음	58	W
12*	**June**	**Alpha**	**높음**	**56**	**B**
13	Mary	Alpha	낮음	54	B
14	Kim	Alpha	낮음	52	W
15	Tom	Alpha	낮음	55	B
16	Ken	Alpha	낮음	48	W
17*	**George**	**Alpha**	**중간**	**48**	**W**
18	Joe	Alpha	낮음	41	B
19	Jeff	Alpha	낮음	44	B
20*	**Heater**	**Alpha**	**낮음**	**37**	**W**

<표17> 교육정도 통계표(원 자료 편집방법 적용) : **노출방지**

county	교 육 정 도				
	합계	낮음	중간	높음	아주 높음
합계	**135**	**50**	**35**	**30**	**20**
Alpha	**20**	13	2	3	2
Beta	**55**	18	12	8	17
Gamma	**25**	5	9	11	0
Delta	**35**	14	12	8	1

비밀편집 방법은 다차원 통계표에 적용하기 쉽고, 또한 비밀노출을 방지할 수 있는 장점이 있는 반면, 비밀노출 방지를 하였는지 외형적으로 알 수 없는 점이 단점이다.

13) * 는 어린이 이름(first name)과 교육정도를 다른 카운티 자료와 교환한 레코드

2. 마이크로 데이터의 익명화 방법

마이크로 데이터에 대하여 비밀노출 위험도를 측정하여 적절하게 방지하는 표준화된 익명화 기술은 없다. 지금까지 알려지거나 연구된 방법들은 다음과 같으며, 이들은 <표4>의 공공이용파일에 적용하는 방법들이다.

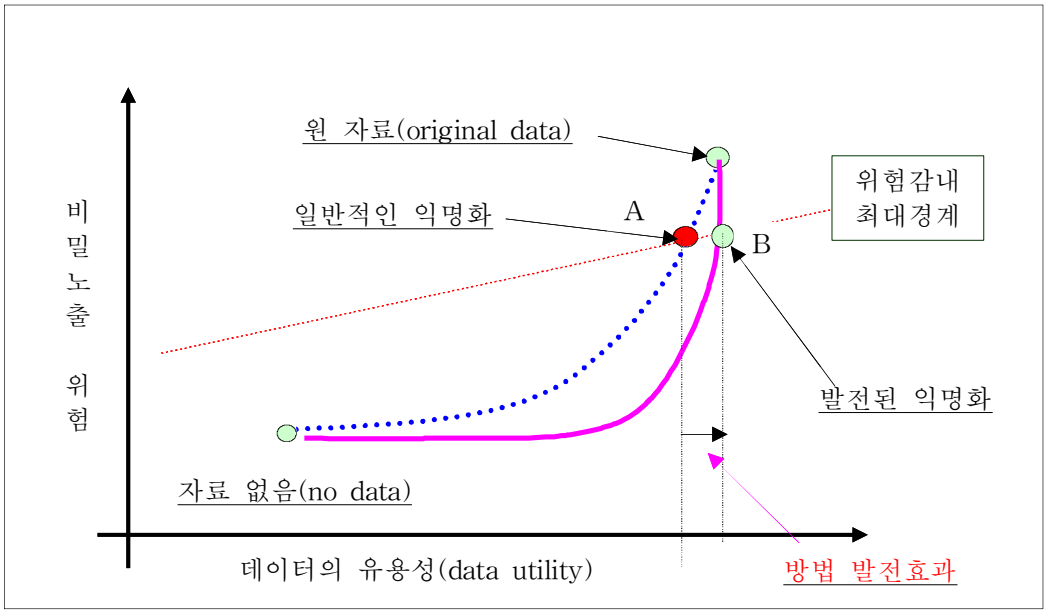
(1) 표본방법(Sampling) : 표본자료로만 마이크로 데이터 파일을 만드는 방법
(2) 식별자(identifier) 제거 : 파일에서 개인자료 식별 항목을 제거하는 방법
(3) 지역적(geographic)으로 세분의 정도를 제한하는 방법
(4) 변수를 제한하는 방법 : 민감한 항목을 제한하는 방법

(5) 상한 또는 하한 코딩(top or bottom-coding) 방법
(6) 재 코딩(recoding)하는 방법 : 간격(interval) 또는 올림(rounding)으로 변환
(7) 잡음(noise) 추가 방법 : 임의 값을 더하거나 곱하는 방법
(8) 자료교환(swapping) 또는 순위교환(rank swapping) 방법

(9) 임의의 레코드를 선택하여 민감한 변수를 공백으로 만든 후에 대체법(imputation)을 적용하여 마이크로 데이터 파일을 만드는 방법
(10) 희석(blurring) 방법 : 응답자 자료를 적당한 그룹으로 집계한 후, 평균하여 그 값을 대체하는 방법

이상과 같이 10가지 방법이 알려진 마이크로 데이터에 대한 익명화 방법이다. 물론 이러한 방법들을 조합하거나 응용한 새로운 방법들도 있다. 이와 같은 데이터의 익명화 방법은 비밀노출 위험방지와 데이터 유용성(utility)이 반비례 관계이므로 적절한 조화를 이루는 방법이 최적의 익명화 방법이다. 앞에서 언급한 익명화 방법들 중에서 (1)~(4)은 데이터의 레코드 단위로 익명화 방법으로 데이터의 유용성보다 익명화에 초점을 둔 방법이며, (5)~(10) 방법은 데이터 유용성을 높이고자 항목내용을 익명화하는 방법으로 알려져 있다. 하지만, 아직 이 방법들에 대한 비밀노출 위험성과 데이터의 유용성과의 관계를 명확하게 측정하는 방법은 없다.

[그림10]은 익명화 방법을 적용함으로써 얻을 수 있는 데이터의 유용성 증가 효과를 그림으로 나타낸 것이다.



[그림10] 익명화 기술적용에 따른 효과 분석

[그림10]의 X축은 데이터의 유용성을 나타내고, Y축은 비밀노출 위험성을 나타내고 있다. 약한 점선은 감내할 수 있는 비밀노출 위험의 정도를 나타는 선이며, 진한 점선은 일반적인 익명화 기술, 실선은 발전된 익명화 기술을 적용한 경우이다. 다시 말하여 “B”와 같은 마이크로 데이터를 이용자에게 공급할 수 있다면, 원 자료(original data)와 거의 같은 수준의 데이터 유용성을 가지고 있기 때문에 가장 이상적일 것이다. 이와 같은 결과를 얻기 위하여 많은 학자들이 노력하고 있다[1][2][13][14][15][16][17].

2.1 표본, 식별자(identifier) 제거, 지역 세분정도 제한 방법

<표15>의 내용과 유사하게 동일 자료에 대하여 위의 3가지 방법을 적용하면, <표18>의 마이크로 데이터 파일을 얻는다. <표 18>에서 알 수 있듯이 지역 세분정도를 조정의 예를 보면, "Alpha" 또는 "Gamma"를 "AlpGam"이라

하여 세분의 정도를 조정하고 있다.

<표18>가상 마이크로 데이터 : 표본, 식별자 제거, 지역세분 제한 방법

번호	county	교육정도	소득	인종
1	AlpGam	높음	61	W
2	AlpGam	낮음	48	W
3	AlpGam	중간	30	B
4	AlpGam	중간	52	W
5	AlpGam	아주 높음	117	W
6	Beta	아주 높음	138	B
7	Beta	아주 높음	103	W
8	Beta	낮음	45	W
9	Beta	중간	62	W
10	Beta	높은	85	W
11	Delta	낮음	33	B
12	Delta	중간	51	B
13	Delta	중간	59	W
14	Delta	낮음	72	B

주) AlpGam은 Alaph 또는 Gamma, 소득 : 천 달러

2.2 상한(top), 하한(bottom) 코딩, 구간 재코딩(recoding into intervals)

<표18>의 가상 데이터에서 “소득” 항목이 포함되어 있다. 이 항목은 가족을 유일하게 식별할 수 있는 항목이 될 수 있어, 이를 민감(sensitive) 항목 또는 높은 시각(high visibility) 항목이라 한다.

이와 같은 민감한 항목은 상한, 하한, 구간 재코딩 방법을 이용하여 비밀노출 위험을 줄일 수 있다. <표18>에 대하여 소득 항목을 상한, 하한, 구간 재 코딩 방법을 적용한 결과가 <표19>이다. 이 표의 상한은 연간 10만 달러 이상을 정하였고, 하한은 40만 달러 미만으로 하였다. 중간의 재 코딩은 1만 달러 간격으로 구간간격을 정하였다. 상하한 구간 및 구간 간격은 보고서 등 이미 작성된 보고서의 기준 또는 작성할 보고서의 통계표 기준과 동일하게 한

다면, 보고서 내용과 차이가 없는 결과를 얻을 수 있다. 또한 <표18>에는 소득에 대한 구간 자료는 실질적인 값으로 표현하고 있지만, 실제 자료에는 코드로 변환하여 활용한다면 효과적일 수 있다. 이 경우에 메타 데이터 관리 및 이용자에 제공해야하는 작업이 필요하다.

<표19>가상 마이크로 데이터 : 소득부분 상한, 하한, 구간 조정

번호	county	교육정도	**소득**	인종
1	AlpGam	높음	60-69	W
2	AlpGam	낮음	40-49	W
3	AlpGam	중간	<40	B
4	AlpGam	중간	50-59	W
5	AlpGam	아주 높음	>100	W
6	Beta	아주 높음	>100	B
7	Beta	아주 높음	>100	W
8	Beta	낮음	40-49	W
9	Beta	중간	60-69	W
10	Beta	높은	80-89	W
11	Delta	낮음	<40	B
12	Delta	중간	50-59	B
13	Delta	중간	50-59	W
14	Delta	낮음	70-79	B

주) AlpGam은 Alaph 또는 Gamma, 소득 : 천 달러

2.3 임의 잡음 추가(adding random noise) 방법

소득과 같이 아주 민감한 항목에 대한 새로운 방법으로 임의의 숫자, 즉 잡음을 더하거나 곱하는 방법이 있다.

이 방법에 대한 예는 <표18>을 이용하면 <표20>과 같다. 예를 위하여 <표18>의 소득 항목에 평균이 0, 표준편차가 5인 정규분포의 임의 변수 값을 추가하는 것으로 가정하였다. 그 결과 <표20>과 같은 결과를 얻었으며, 소득 내용에 잡음이 추가된 부분은 진하게 된 부분으로 거의 모든 레코드에서 이루

어져 있다. 이 방법의 특징은 지정된 평균과 분산의 범위 내에서 잡음이 추가되므로 원 자료의 유용성을 해치지 않는데 있다.

<표20>가상 마이크로 데이터 : 잡음추가 방법 적용 예

번호	county	교육정도	**소득**	인종
1	AlpGam	높음	61	W
2	AlpGam	낮음	**42**	W
3	AlpGam	중간	**32**	B
4	AlpGam	중간	52	W
5	AlpGam	아주 높음	**123**	W
6	Beta	아주 높음	138	B
7	Beta	아주 높음	**94**	W
8	Beta	낮음	**46**	W
9	Beta	중간	**61**	W
10	Beta	높은	**82**	W
11	Delta	낮음	**31**	B
12	Delta	중간	**52**	B
13	Delta	중간	**55**	W
14	Delta	낮음	**61**	B

주) AlpGam은 Alaph 또는 Gamma, 소득 : 천 달러

2.4 교환(swapping) 또는 순위 교환(rank swapping) 방법

교환방법은 추출된 표본 레코드에 대하여 이루어지며, 미리 정해진 변수(항목)들의 집합에 대하여 데이터베이스의 레코드와 연계하여 교환한다. 그 예는 비밀 편집(confidentiality edit) 방법의 일부와 동일하다. 즉 <표15>를 <표16>으로 변화하는 방법이 그 예이다.

순위교환 방법은 교환방법과 유사하지만, 교환대상 항목을 연속적인 쌍으로 구성된다는 점이 다르다. 그 결과 정확하게 일치하는 경우가 거의 없으므로 가장 근접하는 레코드의 항목 쌍을 교환한다. 이를 위하여 관련 항목 쌍은 순서정렬(sort)이 중요한 업무처리 요소가 된다.

2.5 공백(blank)과 대체(impute) 방법

공백과 대체(blank and impute) 방법은 마이크로 데이터 파일로부터 약간의 레코드만 선택한 후, 선택된 항목을 공백으로 바꾼 후에 대체법(imputation)을 적용하여 공백부분을 채우는 방법이다.

예를 들면, <표18>의 카운티 항목에서 “AlpGam", "Beta", "Delta"에 대하여 임의로 한 개의 레코드를 선정한다. 결과는 2번, 6번, 13번 레코드라고 가정하고, 각각에 대한 소득 항목의 값을 63, 52, 49로 대체한 결과가 <표21>이다.

<표21>가상 마이크로 데이터 : 공백과 대체 방법

번호	county	교육정도	소득	인종
1	AlpGam	높음	61	W
2	**AlpGam**	**낮음**	**63**	**W**
3	AlpGam	중간	30	B
4	AlpGam	중간	52	W
5	AlpGam	아주 높음	117	W
6	**Beta**	**아주 높음**	**52**	**B**
7	Beta	아주 높음	103	W
8	Beta	낮음	45	W
9	Beta	중간	62	W
10	Beta	높은	85	W
11	Delta	낮음	33	B
12	Delta	중간	51	B
13	**Delta**	**중간**	**49**	**W**
14	Delta	낮음	72	B

주) AlpGam은 Alaph 또는 Gamma, 소득 : 천 달러

2.6 희석(blurring) 방법

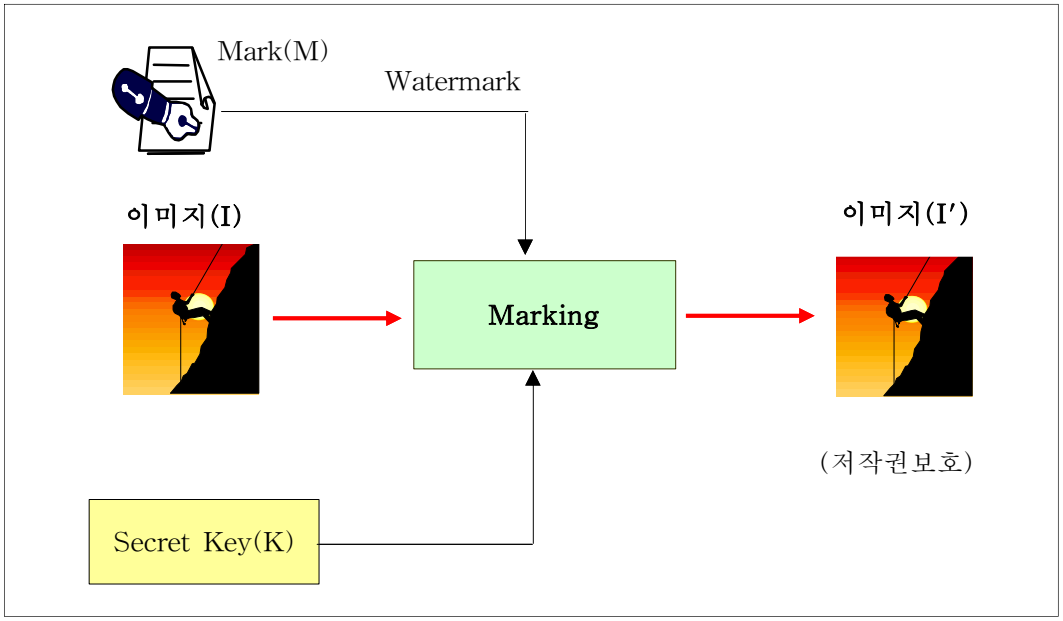
희석방법은 항목 평균값을 항목 값으로 대체시키는 방법으로 이를 구현은 다양한 방법이 있다. 예를 들면, 지정한 항목 값을 임의의 그룹으로 분류한 후, 각 그룹의 평균값을 항목 값으로 대체하는 방법이 대표적이다. 이 방법의 경우도 순서정렬이 중요하다.

3. 무단복제 방지기술 : 디지털 워터마킹(digital watermarking)

최근 인터넷 등 통신망이 발전하면서 디지털 컨텐츠의 저작권 침해가 높아지고 있어 이를 방지하는 기술이 필요하게 되어 **디지털 워터마킹**에 대한 연구가 높아지고 있다.

3.1 디지털 워터마킹, 워터마크의 정의

디지털 컨텐츠에 사용자의 ID나 개인정보를 넣음으로써 불법적인 복제를 제어하고, 데이터 소유자의 저작권과 소유권을 효율적으로 보호하기 위한 방법을 디지털 워터마킹(watermarking)이라 한다. 워터마크(watermark)는 디지털 정보나 기존의 아날로그 정보를 디지털화 할 때, 추가하는 일종의 저작권 관리 정보로서 개인의 식별기호나 부호를 말한다.



[그림11] 워터마킹 개념도

[그림11]은 디지털 워터마크, 워터마킹의 개념을 표현한 그림이다. 그림에

서 이미지(I)가 이미지(I')로 변화하는 과정에서 저작권이 있음을 표시하는 마크 작업이 있고, 이를 해소할 때는 키(key)를 활용한다. 이때 사용하는 핵심적인 기술은 암호화 기술이다.

3.2 디지털 워터마크의 분류

디지털 워터마크는 저작물의 종류, 인지정도, 활용목적 등에 따라 다양하게 분류할 수 있다.

(1) 저작물의 종류에 따른 분류

이미지 워터마킹, 오디오 워터마킹, 비디오 워터마킹, 문자(텍스트) 워터마킹 등이 있다.

(2) 인지 정도에 따른 분류

그림에 자신의 인장을 찍는 것과 같은 실제로 그림을 그린 사람을 알 수 있도록 하는 보이는 워터마킹 방법과 특별한 처리를 하지 않으면 소유자를 알 수 없도록 하는 보이지 않는 워터마킹 방법이 있다.

(3) 활용 목적에 따른 분류

강인한 워터마킹, 연약한 워터마킹, 제거되지 않는 워터마킹 등이 있으며, 일반적으로 말하는 워터마킹은 강인한 워터마킹을 말한다. 강인한 워터마킹은 저작권의 소유자를 식별하는 목적으로 사용하며, 연약한 워터마킹은 약간의 변형으로 마크가 사라지는 방법으로 위조 또는 변조 방지 목적으로 사용한다. 마지막으로 제거되지 않는 워터마킹은 불법적인 유통을 방지하는 목적으로 사용하며, 보이는 워터마킹 기술을 주로 사용한다.

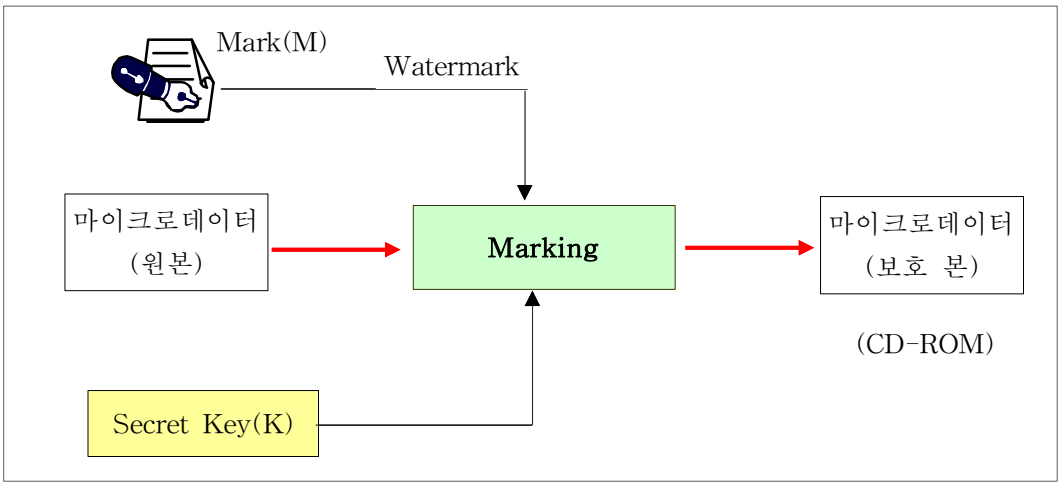
3.3 마이크로 데이터의 불법 복제에 디지털 워터마크의 활용방법

디지털 워터마크는 저작물의 종류, 인지정도, 활용목적 등에 따라 다양한 방법으로 워터마킹 기술을 응용할 수 있다. 통계자료의 마이크로 데이터는 유료

로 판매되는 자료에 대하여 불법복제 방지가 필요하지만, 워터마킹 기술은 원래 이미지 자료, 영상자료 등을 위하여 개발되었고, 또한 최신 기술인 까닭에 마이크로 데이터에 활용하는 방법이나 사례는 거의 알려져 있지 않다. 그러므로 마이크로 데이터에 응용하는 것이 그리 쉽지 않다. 하지만, 마이크로 데이터의 특징과 용도를 알아서 응용한다면 그리 어려운 일이 아닐 것이다. 이 기술을 응용하기 위해서 마이크로 데이터의 특징과 목적을 알아보면 다음과 같다.

(1) 텍스트 데이터(text data)이다.
(2) 인지정도는 통계기관에 소유권이 있다는 것을 알 수 있도록 하여야 한다.
(3) 불법 유통 방지가 필요한 자료이다.
(4) 불법복제 방지의 목적은 유료로 판매되는 CD-ROM 수록자료로 구입자가 구입목적 외에 사용할 수 없도록 하는 것이다.

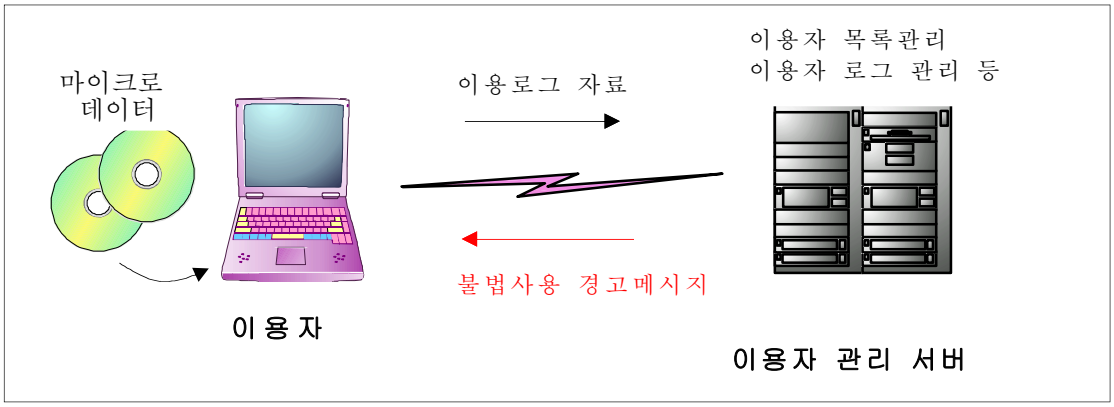
이상과 같은 특징과 목적을 고려하여 시스템을 설계하면 [그림12]와 같은 구성으로 불법복제를 방지할 수 있는 시스템이 될 수 있다. 이를 위한 시스템 개발은 별도로 필요하며, 본 논문에서는 취급하지 않는다.



[그림12] 마이크로 데이터 불법복제 방지 개념

[그림12]의 절차에 의하여 만들어진 마이크로 데이터는 [그림13]과 같은 절차에 의하여 불법복제 CD-ROM의 경우, 즉 위치불명, 사용이 허가된 자가

아닌 경우는 이용자에게 불법사용임을 경고하는 메시지를 전송토록 하여 불법 복제 자료임을 알려준다. 이를 위하여 모든 자료의 사용은 인터넷에 연결되지 않으면 사용할 수 없도록 워터마킹 되어 있어야 한다.



[그림13] 이용자 및 불법사용 관리 절차

4. 다운로드 방지기술 : 원격접속 이용자에 대한 보안대책

다운로드(down-load) 방지는 원격지에서 마이크로 데이터가 저장되어 있는 시스템에 접속하여 사용이 가능한 환경에서 자료의 보호를 위하여 필요한 방법으로 마이크로 데이터를 데이터베이스에 수록하여 외부 이용자들에게 서비스하는 시스템을 설계할 때, 원격지에서 다운로드 불가능하도록 시스템을 설계하여야 한다.

5. 데이터 실험실 운영지침

데이터 실험실은 비밀노출 위험성이 있는 자료에 대하여 외부 이용자가 통계기관의 특정장소에 방문하여 마이크로 데이터를 활용하는 제도이다. 이 제도의 효과를 극대화하기 위해서는 3장 3절의 '2.데이터 실험실'에서 언급한 5가지 구비조건을 철저히 이행할 수 있도록 하는 운영지침이 필요하다.

제 5 장 비밀보호 현황 및 개선 방안

본장에서는 실제적으로 마이크로 데이터 서비스와 이때 사용하고 있는 비밀노출 방지 방법에 대하여 알아보고, 개선방향을 제시한다. 물론 우리나라의 경우 분산형 통계제도이므로 130여 통계작성기관이 각기 다른 마이크로 데이터 서비스 제도와 비밀노출 방지 방법으로 이용자들에게 서비스 하고 있다. 본 연구에서는 통계청의 마이크로 데이터 서비스 형태, 제도, 비밀노출 방지방법 등에 대한 현황을 분석하고, 앞에서 논의한 방법과 비교하여 개선 방향을 제안하고자 한다.

제 1절 마이크로 데이터 서비스 제도

통계청에서 작성하는 52종 통계조사 중에서 30여종 통계의 마이크로 데이터는 이용자들에게 서비스 하고 있다[7]. 그 내용을 보면, 마이크로 데이터를 재집계하가나 마이크로 데이터를 직접 제공하는 등 다양한 형태로 서비스 하고 있다. 이러한 자료의 유형과 서비스 형태는 <표22> 및 <표23>과 같이 분류할 수 있으며, 이는 3장에서 연구한 방법과 거의 일치하고 있다.

<표23>에는 주문에 의해 마이크로 데이터를 재집계하여 제공하는 방법, 마이크로 데이터를 익명화하여 인터넷 쇼핑몰에서 판매하는 방법, 통계목적의 경우는 원 자료를 직접 제공하는 방법, 비밀노출 위험이 높은 마이크로 데이터의 경우는 이용자가 직접 지정 장소에 방문하여 원하는 방법으로 분석하여 그 결과인 매크로 데이터만 외부로 반출하는 제도, 즉 "On Site Access" 등이 있다. 물론 이들 자료는 이용자에게 유료로 서비스 하고 있지만, 국가 및 지방정부는 무료로 보급하고 있다. 민간기관의 경우는 대행기관을 지정하여 재집계 작업 등을 대행하고 있다. 이와 같은 제도의 운영을 위해 관련 지침을 마련하여 운영하고 있다[18][19][20].

“원시자료 제공 규정”에서 정의하고 있는 통계자료의 유형은 6가지로 되어 있으며, 다음 <표22>와 같다.

<표22> 통계자료 제공 규정에 정의한 자료유형

규정의 정의	일반적인 용어	자료 내용
◦ 공개자료	◦ 공개 매크로 데이터	◦ 보고서 등을 통하여 공개된 자료
◦ 제한적 공개자료	◦ 익명화 마이크로 데이터	◦ 개별 식별이 불가능한 마이크로데이터
◦ 조사표수록자료	◦ 원 마이크로 데이터	◦ 개별 식별이 가능한 자료
◦ 명부자료	◦ 조사 대상처 목록자료	◦ 통계조사를 위한 대상처 목록자료
◦ 미공표자료	◦ 미공개 매크로 데이터	◦ 지면부족 등의 이유로 공개를 하지 못한 매크로 데이터
◦ 전산지도자료	◦ 통계조사용 지도 등	◦ 통계조사 및 분석을 위한 지도자료

<표23> 통계청의 마이크로 데이터 서비스 제도 및 내용

서비스 형태		제도 또는 자료 내용	비고
통계표(tables) 형태로 가공하여 대외 서비스하는 방법 (주문제 통계서비스 제도)		◦ 주문통계 서비스 제도 - 미공표 자료, 세분자료 등	유료서비스
익명 마이크로 자료파일 (AMF)	공공이용 파일 (public use file)	◦ 마이크로 데이터 주문 또는 쇼핑몰 구매 - 인구주택총조사 2%표본 등 10여 통계조사	유료판매
	인가파일 (licensed file)	◦ 마이크로 데이터 주문에 의한 서비스 - 비밀노출을 방지한 파일 형태 - 원 자료(original data) ◦ 원 자료의 경우는 이용자의 특별인가가 필요한 경우이며, 통계목적으로만 사용 가능	유료서비스
원격접속제도(RAF)		◦ 메일로 주문 자료를 요청하는 방식 ◦ 마이크로 데이터 직접 접속 방법 개발 중 - 매크로 데이터 : 통계정보시스템(KOSIS)	유료서비스 (유료예상) - 무상
데이터 실험실(DL)		◦ On-Site Access 제도 - 모든 유형의 자료	유료서비스 (개발비무료)

이상과 같이 다양한 마이크로 데이터 서비스 방법이 있지만, 이들 방법은 다음과 같이 네 가지 유형의 통계자료 서비스 제도로 정리할 수 있다.

(1) 마이크로 데이터를 재집계하여 서비스 하는 **주문통계서비스 제도**
(2) 익명화 마이크로 데이터의 **인터넷 쇼핑몰 판매 제도**
(3) 이용자가 임의의 방법으로 집계하여 활용하는 **원격접속 제도**
(4) 이용자가 지정된 장소에서 마이크로 데이터를 임의의 방법으로 재집계하여 매크로 데이터로 변환하여 반출하는 **On-Site Access 제도**

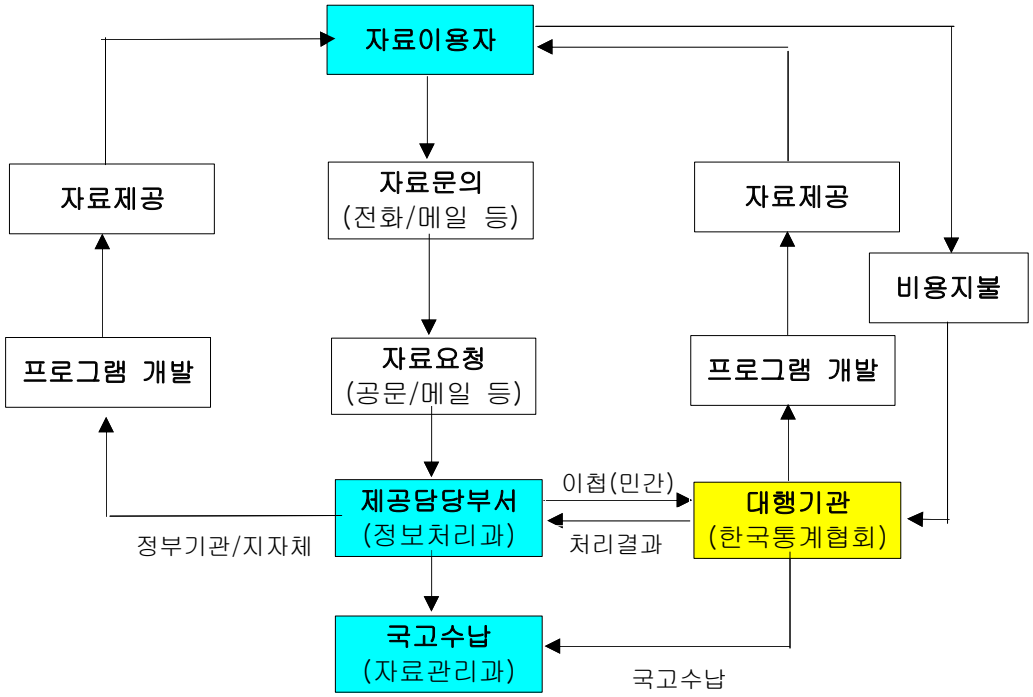
이외에도 통계자료의 대외 서비스 방법, 범위 및 제도개선 등의 심의를 위한 **"통계자료제공심의회"**를 구성·운영하고 있다. 이 심의회에서는 다음과 같은 기능을 수행하고 있다[18].

(1) 통계조사별 자료제공범위의 설정에 관한 사항
(2) 통계목적으로 사용여부의 판단에 관한 사항
(3) 제한적 공개가능 자료에 대한 제공여부에 관한 사항
(4) 자료제공업무의 효율성 향상과 이용자의 편의제고를 위한 제도개선 및 운영방법에 관한 사항 등

통계자료제공심의회의 주요 심의 내용은 주로 자료제공 범위가 모호한 경우에 이를 명확히 하는데 있다. 심의회는 통계기관과 이용자간의 의견 차이를 조정하는 역할을 한다고 볼 수 있다.

1. 주문통계서비스 제도

통계청에서 운영하는 주문통계서비스 제도에서 취급하는 자료유형은 제공 가능한 모든 것을 취급하고 있다. 즉 매크로 데이터, 마이크로 데이터, 명부자료 등 다양하며, 주문 사항을 접수하는 경로도 인터넷, 메일, 팩스, 공문 등 다양하다. 그 자세한 절차는 [그림 14]와 같다.



[그림14] 통계자료제공 처리절차

[표14]에서 보는 바와 같이 자료 이용자는 우선 자료에 대한 문의를 통하여 통계청에서 제공 가능한 자료의 범위, 특성 등에 대한 충분한 사전 지식 습득이 필요하다. 이 과정에서 필요한 자료의 종류, 유형, 비용 등을 알 수 있어 시간과 비용을 최소화 할 수 있다.

최근 주문통계서비스 제도 운영실적은 아래 <표24>와 같으며, 정부기관에서 마이크로 데이터의 요청이 감소하고 있는 추세를 보이고 있다.

<표24> 주문통계서비스 현황

	주문통계서비스 건수(건)				서비스 수입(천원)			
	2000	2001	2002	2003	2000	2001	2002	2003
국가/지방정부	142	139	113	78	-	-	-	-
민간(유료)	259	276	335	302	32,000	28,256	38,082	33,274
합 계	**401**	**415**	**484**	**380**	**32,000**	**28,256**	**38,082**	**33,274**

2. 인터넷 쇼핑몰 판매제도

쇼핑몰 판매제도는 통계청 홈페이지를 통하여 마이크로 데이터 CD-ROM 등을 판매하는 제도이다. 여기서 판매되는 CD-ROM 수록 마이크로 데이터는 이용자의 요청이 많은 경제활동인구조사 등 통계조사 13종을 판매하고 있다. 판매실적은 <표25>와 같으며, 최근에 급속히 증가하고 있는 추세이다.

<표25> 마이크로 데이터 CD-ROM 판매 현황[14)]

	마이크로 데이터 판매건수(건)				판매수입(천원)			
	2000	2001	2002	2003	2000	2001	2002	2003
CD-ROM 판매	40	99	111	156	11,200	6,636	29,150	31,530

3. 원격접속제도

원격접속제도는 데이터베이스에 수록된 마이크로 데이터를 원격지에서 이용자가 직접 처리하여 결과를 출력하는 방법으로 현재 제도 도입을 위하여 필요한 전산시스템을 개발하고 있다[7].

4. On-Site Access 제도

On-Site Access 제도는 외부 유출이 불가능한 자료를 이용하고자 하는 경우 이용자가 지정된 장소에서 집계·분석하여 그 결과를 활용하는 제도이다. 이 제도는 10년여 전부터 시행하고 있지만, 연간 3~4건에 불과할 정도로 그 활용실적이 저조하다. 그 이유는 이용자들이 주로 사용하는 UNIX, NT 등과 같은 전산 시스템이 아닌 IBM 메인프레임(main frame)을 사용해야 하는 불편이 있다. 이러한 문제 해결을 위하여 기존의 마이크로 데이터를 UNIX 계열의 시스템에서 활용할 수 있도록 자료전환 작업을 2002년부터 추진하여 현재 마무리 단계에 있어 그 불편은 곧 해소될 것으로 보인다.

---

14) 쇼핑몰과 전화 등으로 판매실적을 종합하였으며, 쇼핑몰을 통한 판매가 주를 이룸

제 2절 마이크로 데이터의 서비스 범위

통계청에서 작성하는 통계는 <표26>과 같이 2004년 7월 현재 52종이며, 이중에서 16종은 전수조사, 26종은 표본조사, 가공통계 9종, 나머지 1종은 보고통계이다. 이중에서 이용자가 많은 마이크로 데이터는 <표27>과 같이 인구총조사, 주택총조사 등 13종이다. 이들을 중심으로 마이크로 데이터의 제공범위를 분석한다.

<표26> 통계청 작성 통계현황 : 52종 (2004. 7월 현재)

전수조사 (16종)	5년	◦인구총조사 ◦주택총조사 ◦농업총조사 ◦어업총조사 ◦임업총조사 ◦산업총조사 ◦도소매업 및 서비스업총조사
	2년	◦통계활동현황조사
	매년	◦광업·제조업통계조사 ◦건설업통계조사 ◦전국사업체기초통계조사 ◦농어업법인사업체통계조사 ◦환경산업통계조사
	분기	◦전자상거래 기업체통계조사
	매월	◦인구동태조사 ◦사이버쇼핑몰통계조사
표본조사 (26종)	10년	◦국부통계조사
	5년	◦가구소비실태조사 ◦생활시간조사
	3년	◦통계응답실태조사
	2년	◦통계이용실태 및 수요조사
	매년	◦사회통계조사 ◦도소매업통계조사 ◦서비스업통계조사 ◦운수업통계조사 ◦농가경제조사 ◦농업기본통계조사 ◦농산물생산비조사 ◦어업기본통계조사 ◦양곡소비량조사 ◦어가경제조사
	매월	◦광공업동태조사 ◦소비자물가조사 ◦도소매업동태조사 ◦건설수주통계조사 ◦기계수주통계조사 ◦가계조사 ◦경제활동인구조사 ◦건설기성통계조사 ◦서비스업동태조사 ◦소비자전망조사 ◦제조업생산능력 및 가동률조사
가공통계 (9종)	매년	◦한국의 사회지표 ◦지역소득통계 ◦사망원인통계
	매월	◦경기종합지수 ◦설비투자추계지표
	부정기	◦추계인구 ◦생명표 ◦시도별 추계인구 ◦장래가구추계
보고통계(1종)	매월	◦인구이동통계

<표27> 이용도가 높은 마이크로 데이터 종류 및 특징 (2004년 7월 현재)

통계조사	제공범위 및 특징	자료량(MB)	최근년도
◦인구총조사 ◦주택총조사	◦표본 2% 자료 ◦행정구역코드(시군구) 제공	93	2000
◦경제활동인구조사	◦조사대상 가구 전체 제공 ◦조사구 코드 제거 : 전국표본 ◦항목 연계 점검용 항목은 제공하지 않음	100	2003
◦도시가계조사	◦가계부 내용 ◦조사구 코드 제거 : 전국표본	340	2003
◦사회통계조사	◦조사구 코드 제거 : 전국표본	23	2003
◦인구동태통계	◦행정구역코드(읍면동) 제공 ◦출생, 사망, 혼인, 이혼 신고내용	10	2002
◦광공업통계조사	◦행정구역코드(시도 등) 제공 - 산업분류 세분정도에 따라 조정 ◦산업분류 단위로 제공 - 행정구역 분정도에 따라 조정	70	2002
◦사업체기초통계조사	◦행정구역 및 산업분류 : 상동 ◦사업체 명, 매출액 등 제공하지 않음	345	2002
◦가구소비실태조사	◦조사구 코드 제거 : 전국표본	32	2001
◦인구이동통계	◦행정구역코드(읍면동) 제공	160	2000
◦사망원인통계	◦행정구역코드(시군구) 제공	16	2002
◦생활시간조사	◦조사구 코드 제거 : 전국표본	70	1999
◦산업총조사	◦ 광업·제조업통계조사와 동일	71	1998

<표27>에서 보는 바와 같이 마이크로 데이터의 대외 서비스는 거의 모든 자료를 외부 이용자에게 판매하고 있다. 다만, 인구·주택종조사의 경우는 2% 표본에 한하여 외부에 공개하고 있다. 이들 자료들은 표본조사 자료의 경우는 실명화 방지를 위하여 지역 식별 코드(행정구역, 조사구)를 제거하고, 전수조사 자료의 경우는 행정구역코드를 시군구 단위로 가공하여 공개하고 있다.

이상과 같이 마이크로 데이터의 제공 범위에 대한 현황을 살펴보았다.

이들 자료들의 공개범위를 분석하여 비밀노출 방지를 강화함으로써 공개범위를 확대할 수 있다면, 마이크로 데이터의 이용이 활성화되고 또한 공공재인 통계자료의 가치를 제고시킬 수 있을 것이다.

<표27>의 주요 마이크로 데이터 13종 중에서 **전수조사**에 대하여 범위의 적정성을 알아본다. 대표적인 전수조사 통계는 인구·주택총조사와 사업체기초통계조사를 들 수 있다. 이들 통계조사의 공개범위를 비교해 보면, 인구·주택총조사는 2% 표본만 서비스하는 반면, 사업체기초통계조사는 민감 항목인 "매출액"을 제외한 자료 전체를 공개하고 있다. 미국의 경우를 보면, 2000년 인구센서스의 마이크로 데이터는 1%, 3%, 5%로 구분하여 전산자료 파일로 공개하고 있다[7]. 한국의 인구·주택총조사 표본 2% 자료는 공개되는 자료의 양을 조정하는 방안이 필요하다. 예를 들어 2%, 5%, 10%와 같이 차별화하여 이용자가 선택하게 함으로써 자료의 유용성을 높일 필요가 있다. 물론 이때 중요한 것은 범위가 넓어질수록 비밀노출 위험성이 높아지므로 보호방법을 강화해야 하는 등 사전 준비가 필요하다.

경제활동인구조사, 도시가계조사 등 **표본조사** 모든 자료 레코드를 서비스하고 있어 제공 범위 확대는 필요하지 않다. 이들 자료의 경우 특이사항은 전국 표본이므로 행정구역 코드를 삭제하여 오용을 방지하도록 함으로써 자료의 유용성은 부족하지만, 소규모 표본을 지역별 세분하여 분석할 경우에 발생할 수 있는 오용의 위험성을 방지하고 있다.

제 3절 비밀노출 방지 현황

통계청에서 외부에 서비스 하는 통계자료는 통계간행물 및 통계정보기스템(KOSIS) 등을 이용한 매크로 데이터, CD-ROM 등을 이용한 마이크로 데이터가 있다. 본 절에서는 이들 자료에서 비밀노출 방지를 위한 보안방법을 분석하여 문제점을 도출한다.

1. 매크로 데이터의 비밀노출 방지방법

매크로 데이터에 대한 대외 서비스 방법은 여러 가지가 형태가 있지만, 통계조사보고서와 통계정보시스템(KOSIS)이 대표적이다. 이들 자료에서의 비밀노출 방지 방법을 알아본다.

매크로 데이터의 비밀노출 위험은 4장 3절에서 알아본 바와 같이 빈도에 따라 민감 항목이 노출될 위험성이 있다. 따라서 통계청에서는 이를 방지하기 위하여 빈도가 2이하인 경우에는 빈도를 제외한 모든 항목의 통계량을 "X"로 처리하고 있다.

<표28>은 광업·제조업통계조사보고서에서 매크로 데이터의 빈도에 따른 비밀노출을 방지를 위하여 빈도 2이하에 대하여 "X"를 처리한 통계표 예이다. 이 표에는 전국합계, 경상남도의 광업과 제조업, 제주도의 광업 및 제조업을 산업중분류별로 사업체수, 월평균 종사자수, 연간급여액 등을 나타내고 있다.

<표28>에서 "X"처리한 경우를 자세히 살펴보면, 허점을 쉽게 찾을 수 있다. 우선 표의 구조를 보면, 범주속성 부분과 요약속성 부분이 있다. 범주속성 부분은 "C-D 전국", "38 경상남도", "39 제주도"와 같은 지역부분과C "광업", “D 제조업”과 같은 항목부분으로 구성되어 있다. 여기서 유심히 살펴 볼 부분은 이들 범주속성은 다음과 같은 계층구조를 가지고 있으며, 이는 단계별로 집계하는 특징을 가지고 있다.

<표28> 매크로 데이터의 비밀노출 방지방법 : 광업·제조업통계조사보고서[15]

(단위 : 개, 명, 백만원)

분류기호	시도/중분류/연도	사업체수	월평균 종사자 수	연간급여액	생산액	…
C-D	전국	111,025	2,712,310	55,211,123	636,150,600	…
	…					
	…					
38	경상남도 2002	8,067	281,974	6,377,134	61,845,795	…
C	광업	50	760	14,539	122,910	…
C10	석탄, 원유 등 광업	3	25	484	1,455	…
C11	금속광업	-	-	-	-	
	…					
D	제조업	8,026	281,214	6,362,595	61,722,885	…
D15	음·식료품 제조업	731	21,377	314,209	4,243,811	…
D16	담배제조업	2	X	X	X	…
	…					
D37	재생용가공원료생산업	33	301	4,272	41,277	…
39	제주도 2002	329	4,783	70,960	641,285	…
C	광업	11	132	2,539	19,581	…
C10	석탄, 원유 등 광업	1	X	X	X	…
C11	금속광업	-	-	-	-	…
C12	비금속 광물광업	10	123	2,433	19,403	…
D	제조업	318	4,651	68,421	621,704	…
D15	음·식료품 제조업	102	1,919	28,484	321,812	…
D16	담배제조업	-	-	-	-	…
D17	섬유제품제조업	1	X	X	X	…
D18	봉제의복 및 모피제품	3	20	76	296	…
D19	가죽,가방 및 신발	-	-	-	-	…
D20	목재 및 나무제품	8	50	529	3,074	…
D21	펄프, 종이 및 종이	8	183	3,188	28,661	…
D22	출판, 인쇄 및 기록	22	593	9,298	35,396	…
D23	코그스, 석유정제 및	1	X	X	X	…
	…					
D27	1차 금속산업	1	X	X	X	…
	…					
D30	컴퓨터 및 사무용	1	X	X	X	…
D31	기타 전기기계 및	10	67	826	4,984	…
	…					…
D37	재생용가공원료생산업	33	301	4,272	41,277	…

15) 2002광업제조업통계조사보고서 지역편의 Ⅰ-1 시도 산업중분류 및 연도별 통계표 중 일부

(1) 지역계층 : 동·읍·면 → 시·군·구 → 시·도

(2) 항목계층 : 품목분류 → 세세분류→ 세분류 → 소분류→ 중분류→ 대분류

또한 지역계층과 항목계층은 통계표에 따라 병합하여 집계하는 경우도 있다. <표28>의 경우를 보면, “시도” 항목에 대하여 “산업 중분류”와 병합하여 집계하고 있다. 이러한 경우에 **빈도에 따라 "X"로 처리하는 매크로 데이터의 비밀노출 방지는 항목 및 지역의 단계별 집계 특징을 고려하여야 한다**. 즉 하위계층에서 빈도 1, 또는 2가 1개 부분에서 나타날 경우, 상위계층의 합을 이용하면 쉽게 특정한 부분의 비밀사항을 유추할 수 있다. <표28>을 이용하여 빈도를 고려한 비밀노출 방법의 허점의 예는 다음과 같다.

(1) “제주도 광업” 부분 : "C10"의 개별 사업체 내용 노출위험 높음

그 이유는 제주도에는 산업중분류 "C10"의 사업체가 1개 있어 "X"처리하고 있지만, "C10" 전체의 각 항목별 값에서 잔여 산업분류의 합을 제외하고 나면 "C11, 9, 106, 178"이라는 개발 사업체의 정보 노출이 가능하다. 이와 같이 단계별 집계일 경우에는 빈도 1이하 부분이 1개일 경우에는 위험성이 높은 문제점이 있다.

(2) “제주도 제조업” 부분 : "D17, 23, 27, 30," 개별 내용 노출위험 없음

제조업 부분의 경우는 빈도 1인 산업중분류가 4개가 있어 제조업 전체 합에서 개별 사업체별 자료를 찾아내는 방법은 쉽지 않다.

(3) “경상남도 제조업” 부분 : "D16"의 개별 사업체 내용 노출위험 있음

경상남도의 제도업 부분의 경우는 "D16" 부분에 빈도가 2 이므로 "X"처리하고 있지만, 제조업 전체의 합에서 잔여 산업분류의 합을 제외하면 2개 사업체의 합을 찾을 수 있어 경쟁업체의 비밀사항을 쉽게 알 수 있다.

이상과 같이 빈도에 따라 2이하를 "X"로 처리하여 보안을 한다고 하더라도 단계별 집계의 경우 예와 같은 위험성이 존재하는 문제점이 있다. 이러한 문제점을 해결하는 방안을 연구가 필요하다.

2. 마이크로 데이터의 비밀노출 방지방법

마이크로 데이터의 비밀노출 방지를 위하여 <표27>에서 보는 바와 같이 표본방법, 지역 식별 코드를 이용한 방법, 민감 항목을 제외하는 방법 등을 사용하여 익명화하고 있다.

(1) 표본방법 : 인구·주택총조사 2% 표본

표본방법은 조사자료 중에서 일부 표본을 추출하여 마이크로 데이터 파일을 생성하고, 또한 행정구역코드를 읍면동에서 시군구로 변경시켜 익명화하여 인구·주택총조사의 서비스용 마이크로 데이터 파일 작성에 활용하고 있다. 인구·주택총조사는 전수 90%와 표본 10%로 나누어 조사하고, 표본 10% 중에서 20%를 추출하여 전체의 2%에 해당하는 2% 표본 자료를 생성하여 쇼핑몰에서 판매하고 있다.

(2) 지역 식별 코드를 이용하는 방법 : 경제활동인구조사 등

지역 식별 코드는 계층적 구조이므로 이를 상위 그룹으로 변경함으로써 지역적 식별성을 줄이는 방법으로 적합하다. 표본조사인 경제활동인구조사의 경우 실제 자료는 조사구 단위로 조사한 자료이지만, 대외 서비스용에서는 식별이 불가능 하도록 조사구 코드를 삭제하고 있다. 그 결과 전국 기준의 분석만 가능하고, 지역적인 분석은 불가능하도록 하여 비밀노출 및 표본추출의 목적외의 사용이 불가능하도록 하고 있다. 광업·제조업통계조사의 경우는 행정구역코드와 산업분류코들 적절하게 조합하여 비밀노출을 방지하고 있다. 예를 들어 행정구역코드가 읍면동 기준이면 산업분류는 대분류로 하고, 반대로 행정구역코드가 시도 기준이면, 산업분류는 소분류로 하여 식별성을 조정하는 방법이다.

(3) 민감 항목 제거 방법 : 사업체기초통계조사

사업체 부분의 대표적인 전수조사인 사업체기초통계조사의 경우 두 가지 방법으로 마이크로 데이터를 생성한다. 행정구역코드와 산업분류코드는 광업·

제조업통계조사와 동일한 방법을 적용하지만, 매출액과 같은 민감 항목은 삭제하여 마이크로 데이터를 작성한다. 매출액 자료를 삭제하는 것은 민감 항목이기도 하지만, 자료자체의 신뢰도 부족이 주요 이유이다. 신뢰도 부족은 응답자가 매출액 등과 같은 민감 항목에 대한 올바른 응답을 기대하기 어려점도 있지만, 조사목적과 직접 관련이 있는 항목이 아닌 점도 있다.

이상과 같이 현재 마이크로 데이터의 익명화 방법으로 3가지를 주로 사용하고 있다. 물론 이 세 가지 방법만을 사용한다고 명시적으로 정해져 있는 것은 아니다. 이 방법들은 식별성을 낮추어 비밀노출을 방지하는데 목적을 둔 방법이다. 그러므로 데이터 링케이지 문제는 고려하지 않은 단점이 있으며, 또한 임의 방법으로 집계한 결과 빈도 3이하 발생여부의 점검 절차와 발생시 대책에 부족한 점이 있다. 또한 사업체기초통계조사와 같이 일부 항목을 완전히 공개하지 않음으로써 데이터의 유용성을 저하시키고 있다. 따라서 이러한 단점을 보완하는 것이 과제이다.

3. 현행 비밀노출 방지방법에서 보완이 필요한 사항

이상과 같이 매크로 및 마이크로 데이터의 현행 비밀노출 방지 방법에 대하여 알아보았다. 그 알아 본 방법들의 단점을 종합하여 정리하면 다음과 같다.

(1) 매크로 데이터 경우 빈도에 따른 **"X"처리 방법 보완**이 필요하다.

(2) 마이크로 데이터의 경우 실명화 가능성은 거의 없어 보이지만, **데이터 링케이지 방법으로 실명화 위험성 점검**과 발생시에 대책이 마련되어 있지 않다.

(3) 생성된 마이크로 데이터에 대하여 임의의 방법으로 집계하여 **빈도 3이하 발생 여부 점검과 발생시의 처리방법**이 없다.

(4) **데이터의 유용성 증대를 위하여 필요한 비밀노출 방지방법을 적용**하여 가능한 미공개 항목을 공개하려는 노력이 필요하다.

제 4절 비밀노출 방지를 방법 제안

본 절에서는 지금까지 알아 본 현행 매크로 및 마이크로 데이터 비밀노출 방지 방법의 약점을 보완하는 방안에 대하여 제안한다. 제안하는 내용을 실제 정책으로 도입하기 위해서는 관련 전산 시스템 수정 등 추가적인 작업이 필요할 것이다. 이를 위한 부분은 본 연구의 범위에서 제외한다.

1. 매크로 데이터의 빈도에 따른 "X" 방법의 보완 방안

문제발생 원인을 보면, <표28>에서 나타난 예에서 보듯이 빈도가 "1" 또는 "2"가 나타나는 횟수가 "2"이하 일 때 해당 집계 상위 계층의 합을 이용함으로써 비밀사항이 노출되는 문제가 있다. 이의 해결 방안으로 다음과 같은 방법을 생각할 수 있다.

(1) 빈도 "1", "2"(또는 "3")의 경우가 해당 집계그룹에서 1회일 경우는 항목 값들은 물론 해당 빈도까지 "X"로 표현하는 방법

이 방법은 기존의 운영중인 전산시스템의 수정이 쉬워서 당장 적용할 수 있는 장점이 있는 반면, 빈도 "1"인 경우에는 아직도 허점이 남아 있다. 예를 들어 <표28>의 "제주도 광업" "C10"을 모두 "X"로 하여도 다음과 같은 문제가 발생한다.

{"C" 값 - "C11" 값} ={1, 9, 106, 178}= "C10"의 값

이와 같이 "C10"의 값의 유추가 가능하다. 그러나 빈도 2이상의 경우는 그러한 문제가 없으므로 빈도를 "X"하기 전보다는 보안성이 높은 방법이다.

(2) 빈도 “2”이하(또는 “3”)의 경우가 해당 집계그룹에서 1회일 경우는 항목 값들은 물론 해당 빈도까지 "X"로 표현하는 방법

빈도 2이하, 즉 0, 1, 2 모두를 "X"로 표현하는 방법으로 빈도 0은 여러 곳에서 나타날 경우에는 앞의 (1)에서의 문제는 해결할 수 있다. 또한 기존의

프로그램 수정도 용이한 방법이다. 즉 빈도가 0인 부분을 "X"로 표현하기 때문에 어떠한 범주에서 빈도가 0, 1, 2인지 알 수 없도록 하는 방법이므로 보안성과 유용성을 동시에 확보할 수 있는 방법이지만, 빈도 0인 부분이 많으면 "X"로 된 부분이 많아서 외형적인 형식이 복잡하게 보이는 단점이 있다.

(3) 빈도 "1", "2"(또는 "3")의 경우가 해당 집계그룹에서 1회일 경우는 항목 값들만 "X"로 표현하되, 원 자료에 잡음(noise)을 추가하여 재집계하는 방법

빈도 1, 또는 2가 1회만 나타나는 경우에 빈도와 항목 값들을 변경시키는 방법이다. 예들 들어 빈도가 실제 1이지만 2로 변경하고, 관련 항목의 값도 변경함으로써 개별 정보의 유추가 불가능 하도록 한다. 이 방법은 4장 3절의 셀 값 감추기 방법(suppression) 방법과 비밀편집(confidentiality edit) 방법 등을 응용한 방법이라 할 수 있다. 이 방법은 개별정보 유출은 방지할 수 있으나, 실제 적용하기 위해서는 관련 프로그램 알고리즘이 복잡하여 적용하는데 많은 노력이 필요한 단점이 있다.

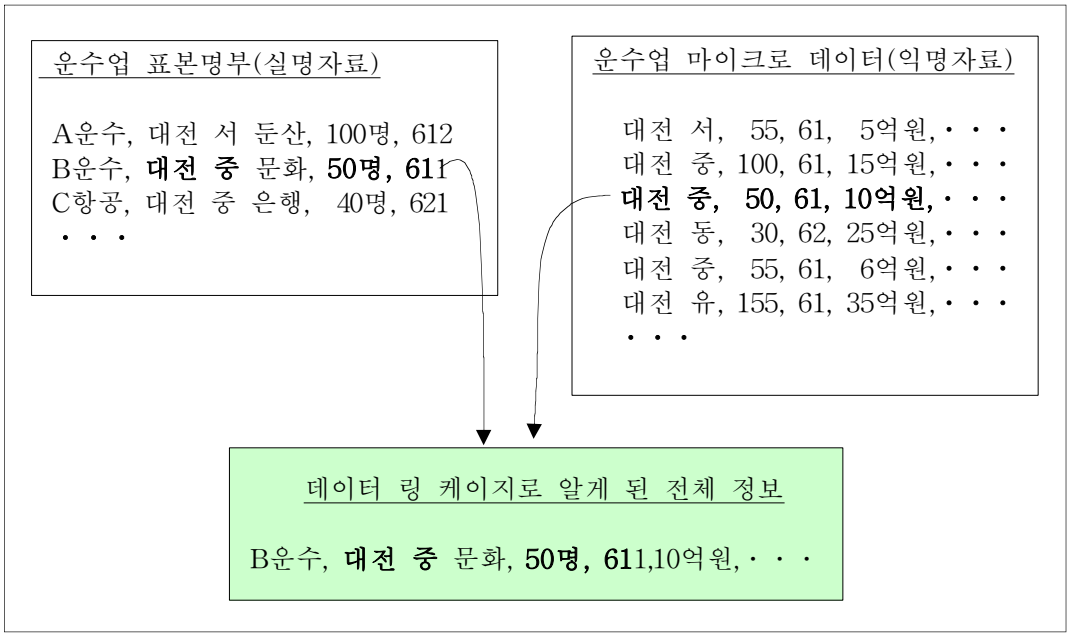
이상과 같이 매크로 데이터의 비밀노출 방지 방법의 보완 방안 3가지를 제시하였다. 이 들 방법 외에도 **4장 3절의 방법들을 응용하면 여러 가지 방법**을 생각할 수 있다. 하지만, 가장 중요한 것은 해당 통계조사 자료의 특성과 보안의 요구 정도에 따라 적합한 방법을 선택하여야 한다. 이때 고려해야 할 사항은 비밀노출 방지를 강화하면 할수록 데이터의 유용성은 낮아지고, 활용성도 떨어지는 것이 일반적인 특징이다. 그러므로 적합한 수준의 보안을 통하여 데이터 유용성의 손실을 최소화 하는 범위에서 해결 방법을 찾아야 한다.

2. 데이터 링케이지 방법에 의한 비밀노출 위험성 점검

데이터 링 케이지 방법은 마이크로 데이터 셋들(micro data sets)을 연계하여 항목내용이 일치하는 경우에 다른 항목의 내용이 알려지는 경우가 있다. 현재 통계청에서 제공하고 있는 마이크로 데이터는 행정구역코드나 조사구 코드와 같은 지역코드가 세분되어 있지 않고, 또한 민감 항목도 서비스 하지 않

고 있어 위험성은 거의 없다.

그러나 조사대상처 표본 추출과 같이 통계조사 목적으로 명부자료를 제공할 경우가 있다. 이때 당연히 매출액, 가계소득과 같은 민감 항목은 명부자료에 포함되지 않으나 동일한 조사 내용의 마이크로 데이터를 제공하는 경우가 있을 수 있다. 이러한 경우에 데이터 링케이지를 이용하여 해당 데이터 레코드 전체 항목이 노출될 위험성이 있다. 예를 들어 운수업통계조사 모집단에 운수업 표본을 추출하여 200개의 사업체 명, 종사자 수, 주소, 산업분류 항목을 제공하고, 한편으로 운수업통계조사의 마이크로 데이터를 제공하였다면, 이들 자료를 연계하면 [그림15]와 같이 상세한 개별정보 확보가 가능하여 진다.



[그림15] 데이터 링케이지를 이용한 개별정보 노출의 예

[그림15]에서 명부자료는 민감 항목이 없지만, 운수업의 마이크로 데이터에는 매출액과 같은 민감 항목이 있다. 이들 항목을 상호 연계, 진한 글씨 부분의 값들을 연계하면 “B운수” 업체의 주소, 종사자 수, 산업분류, 매출액 등 전체 레코드 정보의 확보가 가능하여 진다. 따라서 이러한 문제의 해결을 위한 방법이 필요하다.

(1) 임의의 항목에 대한 최고 또는 최저 코딩(top or bottom coding) 방법

표본명부에서 종사자 수를 조사의 특성을 고려하여 적정한 간격으로 구간을 설정하는 최고/최저 구간으로 바꾸어 제공하는 방법이 있다. 마찬가지로 마이크로 데이터에도 동일한 방법을 적용할 수 있다. 이렇게 함으로써 상호 연계가 일어나지 않도록 하여 비밀노출을 방지 할 수 있다.

(2) 희석(blurring) 방법 적용

희석방법은 지정된 항목을 임의의 그룹으로 나눈 후에 이들 그룹들의 평균값으로 대체하는 방법이다. 이 방법도 최고/최저 코딩방법과 같은 효과가 있다. 그러나 전산시스템에 적용할 경우에는 희석방법이 최고/최저 코딩 방법보다 복잡한 것이 단점이라 할 수 있다.

(3) 잡음(noise) 추가 방법, 자료교환(swapping) 방법 등 기타

(1)과 (2) 방법 외에도 4장 3절에서 논의한 잡음(noise) 추가 방법, 자료교환 방법(swapping), 공백 후 대체법(imputation) 등 다양한 방법을 고려할 수 있지만, 알고리즘이 복잡하고 데이터의 유용성을 해칠 우려성이 높은 방법들이다. 하지만, 이용자 요구 특성 등에 따라 필요한 경우에 활용할 수 있다.

3. 서비스용 마이크로 데이터 집계 시 빈도 3 이하 발생 여부 점검 및 대책

전수조사 자료의 경우, 서비스용 마이크로 데이터를 집계하면 일부 집계 단위에서 빈도 3(또는 5이하)이 있다면, 매크로 데이터의 빈도에 따른 비밀노출 위험성이 있다. 그러므로 서비스용 마이크로 데이터라 하다라도 서비스 전에 집계하여 빈도가 3이하일 경우에는 별도의 대책이 필요하다. 이를 위한 대책으로 다음과 같은 방법들이 있다.

(1) 원 자료 에디팅을 통한 잡음 추가 방법

마이크로 데이터를 집계함으로써 빈도 3이하라는 것을 알게 되기 때문에 원자료의 에디팅을 거치지 않고는 비밀노출 문제를 해결할 수 없다. 따라서

원자료를 에디팅하여 빈도 3이하가 나타나지 않도록 잡음(noise)을 추가한다. 잡음을 추가하는 방법의 실제 구현은 다양한 방법들이 있을 수 있다. 예들 들어 "C11"이라는 산업분류에서 빈도가 2가 나타났다면, 다른 이웃하는 산업분류의 값을 이용할 수도 있다. 또 다른 방법으로는 "C11"이라는 산업분류가 나타나지 않도록 산업분류를 이웃 산업분류로 변경하는 방법 등이 있을 수 있다.

(2) 기타 방법

4장 3절의 마이크로 데이터의 익명화 방법 중에서 재 코딩 방법, 자료교환 방법, 흐림 방법 등 여러 가지를 적용할 수 있다. 이러한 방법들을 적용할 때는 어떤 방법이 이용자 요구에 적합한 데이터 유용성을 제공할 수 있는지를 고려하여 선택하여야 한다.

4. 미공개 자료의 공개를 위한 비밀노출 방지 방법

마이크로 데이터는 [그림7]과 [그림10]에서 보는 바와 같이 비밀노출 위험(disclosure risk)과 데이터의 유용성(data utility)은 서로 비례적인 관계이다. 그러나 데이터 공급자는 위험성을 낮추어서 응답자를 보호해야 하고, 반면 이용자는 유용성을 높여서 데이터의 활용 가치를 높이고 싶어 한다. 이렇게 서로 다른 목적으로 인하여 공급자와 수요자는 서로 긴장관계에 있다고 볼 수 있다.

통계청에서 서비스 중인 마이크로 데이터 중에서 민감한 일부 항목은 공개하지 않고 있다. 하지만, 여건이 변하여 자의 또는 타의로 공개해야 할 경우가 발생할 수 있다. 이는 최근 정부기관에서 보유하고 있는 정보의 공개를 확대하고 있는 추세이기 때문에 머지않아 피할 수 없는 현실로 다가올 것이다. 이를 대비하여 비밀노출 등 데이터 보호 방법이 필요하다.

(1) 흐림 방법(blurring) 활용 : 매출액 등 민감 항목에 적합한 방법

흐림 방법은 지정항목의 응답자 값을 적당한 그룹으로 분류하여 각 그룹별

평균값을 항목 값으로 대체하는 방법이다. 이 방법의 특징은 그룹을 몇 단계로 설정하는가에 따라 데이터의 유용성이 달라진다. 예를 들어 20개 그룹으로 나누는 경우와 10개 그룹으로 나누는 경우 데이터의 가치가 달라진다. 다시 말하면, 그룹의 수를 줄이면 비밀보호는 강해지고, 수를 늘리면 데이터의 유용성 높아진다.

이와 같은 특징을 이용하여 적정한 보호와 유용성을 확보할 수 있도록 하여 공개하고 있지 않는 자료항목에 대해 서비스 할 준비를 하여야 한다. 사업체기초통계조사의 매출액과 같은 항목이 좋은 사례가 될 수 있다.

(2) 구간(interval)을 이용한 재 코딩(recoding) 방법

재코딩 방법은 지정된 항목의 값을 지정된 수의 구간으로 분할 한 후, 각 구간의 경계를 기존의 값으로 대체시키는 방법이다. 예를 들어 매출액이 최저 500만원부터 10억원까지 있다면, 이를 10개 구간으로 나누어 그 구간에 해당하는 값으로 대체시키는 방법으로 <표19>와 같이 “5000 <”, “5000-8000” 로 표기한다. 이 방법은 흐림 방법보다는 전산시스템에 구현하기 편리한 장점은 있으나, 이용자가 활용할 때는 별도의 코딩작업이 필요한 단점이 있다.

5. 비밀노출 방지 내용을 설명하는 메타 데이터

앞에서 4가지 분야에 대하여 비밀노출 방지 방법에 대하여 활용을 제안하였다. 이들 보안 방법을 활용한 자료를 이용자에게 제공할 경우에 추가적으로 꼭 필요한 사항은 **자료에 적용한 보안 사항 즉 메타 데이터**를 데이터와 함께 제공 해야 한다. 그렇지 않을 경우에 보고서 등 기 발표된 매크로 데이터와의 차이가 발생할 수 있어 데이터의 신뢰도 문제가 제기될 수 있다. 뿐만 아니라 자료 활용에 필요한 각종 코드 북, 파일 설계도 등 메타 데이터도 동시에 제공되어야 하는 것은 당연하다.

제 6 장 결론 및 향후과제

지금까지 통계자료의 이용 활성화를 위하여 통계자료의 특성과 비밀노출 방지 방법에 대하여도 알아보았다.

통계자료는 기획, 현장조사, 자료처리, 집계분석, 서비스 등의 과정을 거쳐 작성되며, 형태에 따라 마이크로 데이터와 매크로 데이터가 구분한다. 마이크로 데이터는 응답자가 응답한 내용을 입력, 오류정정 등의 작업을 완료한 자료이며, 매크로 데이터는 마이크로 자료를 집계한 통계표(table)를 말한다. 또한 통계자료는 요약속성, 범주속성, 메타 데이터로 구성되어 있고, 이 속성으로 인하여 여러 가지 특징이 나타나고 있다. 범주속성의 특징으로 요약속성 자료가 다차원 구조를 이루고, 통계자료의 이용가치를 향상시키는데 메타 데이터의 역할이 중요하다는 것을 알았다.

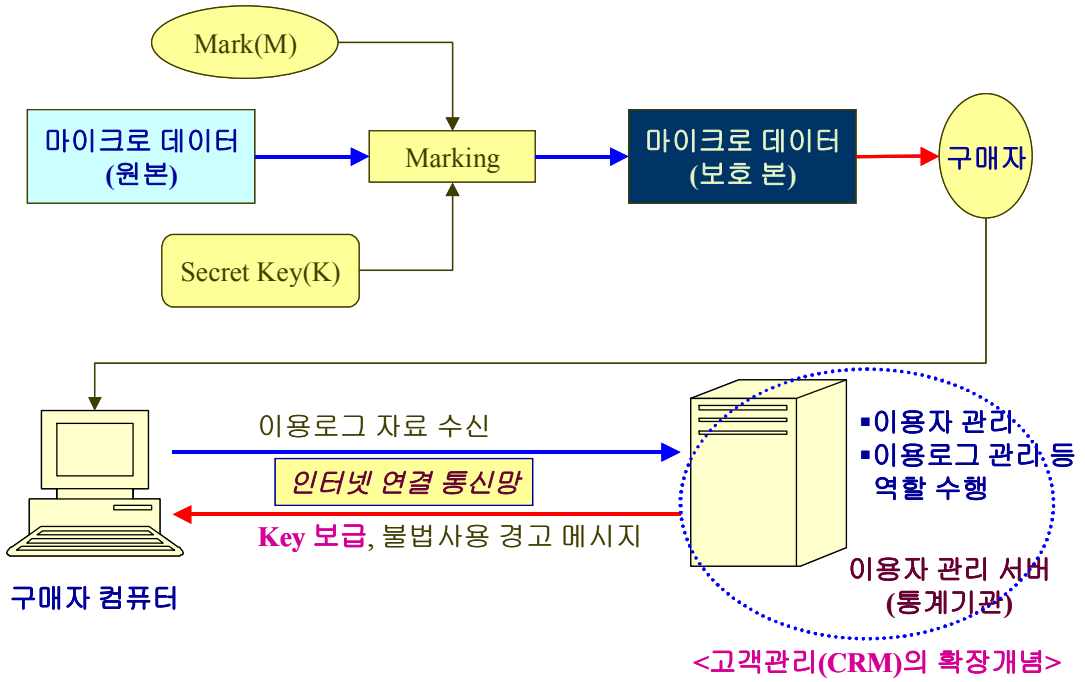
비밀노출 방지에 대해서는 매크로 데이터와 마이크로 데이터에 대한 위험분석과 함께 그의 방지방법에 대하여 알아보았다. 매크로 데이터에는 빈도에 따라 실명화 위험이 있을 수 있어 이를 대비하는 방법, 즉 셀 감추기 방법(suppression), 랜덤 올림 방법(random rounding), 제어 올림 방법(controlled rounding), 비밀편집 방법(confidentiality edit), 마이크로 데이터에 대해서는 표본방법(sampling), 식별자(identifier) 제거 방법 등 10가지 익명화 방법을 예를 통하여 소개하였다. 또한 통계청의 매크로 및 마이크로 데이터의 비밀보호 방법을 분석하여 개선방안을 제안하였다.

통계자료의 비밀보호는 데이터 유용성(data utility)과 실명화 위험성 모두를 만족해야하는 어려움이 있다. 그러므로 실제 매크로 및 마이크로 데이터에 제안한 방법의 적용을 위해서는 적용대상 통계의 특성과 이용자 유형을 충분히 분석한 후 적절한 방법을 선택하여야 한다. 다시 말하여 제안한 방법들이 절대적인 방법이 될 수 없으며, 대상 통계에 따라 다양하게 응용할 수 있다.

이와 같이 통계자료의 이용 활성화를 위해서는 메타 데이터의 체계적인 관리와 적합한 비밀보호 방법의 적용이 필요하다. 이와 관련하여 향후에 연구개발 또는 검토하여야 할 과제들은 다음과 같다.

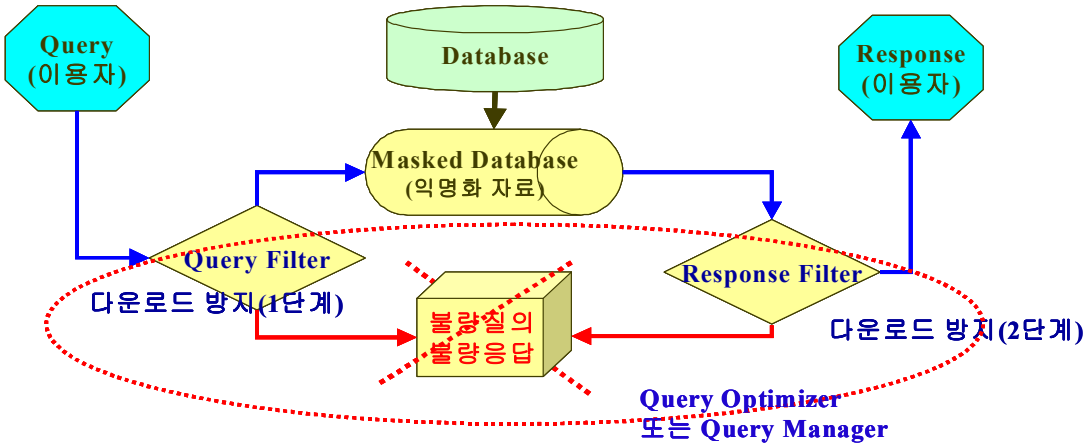
첫째로는 메타 데이터의 체계적인 관리를 위하여 표준화 및 데이터베이스 구축 등의 작업이 필요하다. 이 작업은 광범위하고, 다양하여 단시간에 이루어 질 수 없는 작업이므로 장기적인 연구개발이 필요한 과제이다.

둘째로는 유료로 판매하는 마이크로 데이터의 불법복제 방지를 위하여 워터마킹(watermarking) 기술 도입과 이용자 관리시스템 구축이 필요하다. [그림16]은 마이크로 데이터의 불법복제 방지 체계도 이다. 여기에는 마이크로 데이터의 원본을 보호 본으로 바꾸는 작업에는 워터마킹 기술을 활용하고, 구매자는 반드시 인터넷에 연결된 PC에서만 활용하도록 하여 사용이력(로그자료)을 자동으로 이용자 시스템에 수집되도록 하는 방식이다. 수집된 로그 자료는 향후에 고객관리 체계(CRM)로 확장할 수 있다.



[그림16] 마이크로 데이터의 불법복제 방지 방법

셋째로는 현재 개발 중에 있는 원격접속방법의 마이크로 데이터의 온라인 이용 시스템을 구축할 때 제안한 방법 또는 응용방법을 적용하는 것과 온라인상에서 직접 마이크로 데이터의 다운로드를 방지하도록 시스템을 구현하여야 한다. 이를 위한 시스템 구조는 [그림17]과 같은 형태를 취하면 다운로드 방지 및 불법적인 질의 처리를 방지할 수 있다. [그림17]에서 데이터베이스는 두 종류, 즉 원 자료와 마스킹 자료가 있으며, 원격지에 서비스하는 데이터베이스는 마스킹 데이터베이스이다. 불량 질의와 응답은 질의 투과기(query filter)와 응답 투과기(response filter)에서 식별하여 통제하는 절차로 이루어져 있다.



[그림17] 불량 질의 및 응답 관리 방법

마지막으로 통계기관에서 통계자료 보급을 **위험회피 전략에서 위험관리로 전략으로 전환**하여야 한다. 이는 통계자료의 보급범위의 확대와 관련이 있다. 즉 적극적으로 보급범위는 확대하되, 문제가 되는 부분은 기술적 방법으로 관리하는 것이 공공재인 통계자료 이용 활성화의 가장 핵심 요소라 할 수 있다. 따라서 현행 통계자료의 보급범위 검토하여 부족한 부분은 적극적인 확대가 필요하다. 예를 들어 인구주택총조사의 표본규모를 2%, 5%와 같이 확대하여 통계제품을 다양화 하는 방법, 또는 사업체기초통계조사의 매출액을 익명화 방법을 적용하여 공개하는 방법 등이 될 수 있다.

참 고 문 헌

[1] Federal Committee on Statistical Methodology, *Statistical Policy Working Paper 22. Report on Statistical Disclosure Limitation Methodology*, May, 1994.

[2] George Duncan, "An Overview of Policy and Practice on Release of Micro-Data," The Heinz School, Carnegie Mellon University, Sept. 2002.

[3] Federal Committee on Statistical Methodology, *Statistical Policy Working Paper 24. Electronic Dissemination of Statistical Data*, Nov. 1995.

[4] Dennis Trewin, *Task Force on Confidentiality and Microdata-Discussion Paper*, Statistical Commission and Economic Commission for Europe Conference of European Statisticians. June, 2004.

[5] 박원환, "상용 데이터베이스와 통계 데이터베이스," 통계, 제15원 1호, 1989.

[6] Michael Colledge, Fred Wensing, and Eden Brinkley, "Integrating Metadata with Survey Development in a CAI Environment," Australian Bureau of Statistics.

[7] 통계청, 원시통계자료제공 개선계획, 2004. 4월

[8] Australian Bureau of Statistics, "Integrated Censuses," Year Book, Chapter 31, Catalogue No1300.0, Australian Bureau of Statistics, Belconnen, ACT2616, Australia, 1969.

[9] Bethlehem J.G,, Homan L., Schuerhoff M.H., *Braise Ⅲ, Overview Central Bureau of Statistics,* Voorburg, Netherlands, 1994.

[10] Colledge M.J., "Business Survey Redesign Project : Implementation of New Strategy at Statistics Canada," Proceedings, Third Annual Research Conference, U.S. Bureau of the Census, Washington DC, U.S.A., 1986.

[11] Sundgren B.,"Towards a Unified Data and Metadata System a The Australian Bureau of Statistics," Working Paper, Australian Bureau of Statistics, Belconnen, ACT2616, Australia. 1991.

[12] National Center for Education Statistics, *NCES Statistical Standards,* National Center for Education Statistics, U.S. Department of Education, 2003.

[13] Sandra Rowland and Laura Zayatz, "Disclosure Limitation for American FactFinder," Feb. 1999.

[14] T. Evans, L. Zayatz, and J. Slanta, "Using Noise for Disclosure Limitation of Establishment Tabular Data," Bureau of the Census.

[15] Richard A. Moore, Jr.,"Controlled Data-Swapping Techniques for Masking Public Use Microdata Sets," Statistical Research Division, US Bureau of the Census.

[16] Anna Manning and Mark Elliot, "Applying Disclosure Control to Temporal Data," Dept. of Computer Science, University of Manchester, UK

[17] Meena Khare, Michael P. Battaglia, and David C. Hoaglin, Procedures to Reduce the Risk of Respondent Disclosure in a Public-Use Data File: The National Immunization Survey," National Center for Health Statistics, USA.

[18] 통계청 통계정보국, *원시자료 제공규정*, 1996

[19] 통계청 통계정보국, *대행기관의 통계자료제공업무 처리지침*,1998.

[20] 통계청, 통계정보국, *ON-SITE ACCESS 운영지침*,1998.