국가 통계표준 메타데이터 설계에 관한 연구*

Construction of the Guidelines for National Statistical Metadata

남 영 준**
Young-Joon Nam

차 례

초 록

본 연구에서는 인터넷 상에서 국가 통계 데이터의 자유로운 활용과 공개를 위해 필요한 메타데이터 표준안을 제안하였다. 설계는 우리나라 통계조사보고서와 통계청 내부 확장 메타데이터 기준을 기반으로 국제기준에서 요구하는 최소한의 메타데이터 세트가 포함되도록 하였다. 그 결과, SDMX와 SDDS의 중복요소에서 29개의 항목을 채택하고, 더블린 코어에서 14개의 항목을 채택하여 최종적으로 43개로 이루어진 국가 통계표준 메타데이터를 완성하였다.

키 워 드

통계 데이터, 통계 메타데이터, 국가 통계표준 메타데이터, SDMX, SDDS

* 이 논문은 2004년도 중앙대학교 학술연구비(일반연구비) 지원에 의한 것임.
** 중앙대학교 문과대학 문헌정보학과 교수
(Associate Professor. Dept. of Library and Information Science, Chung-Ang University, namyj@cau.ac.kr)
• 논문접수일자 : 2005년 2월 28일
• 게재확정일자 : 2005년 3월 10일

ABSTRACT

This paper proposes some guidelines for Korean standards for statistical metadata to share on the Internet. The construction includes the minimum metadata set needed in the international standards based on the report and Korean Statistics Office. In result, 29 categories were selected from the SDMX and SDDS' consistency items and 14 categories from Dublin Core. Overall, 43 international statistics metadata was completed.

KEYWORDS

statistical data, statistical metadata, SDMX, Statistical Data and Metadata exchange, SDDS

1. 서 론

통계는 사실적 정의나 개념규정에 기초하여 현실사회의 현상을 표현한 수치 데이터이다. 또한 통계조사는 어떤 목적에서, 구체적으로 무엇을 조사하고, 무엇을 통계로서 표시하는가를 조사하는 행위이다. 따라서 통계조사업무는 고도의 계획에 따라 체계적인 수집과 분석이 필요한 업무로서, 이를 생성하기 위해서는 많은 예산과 인력을 필요로 한다. 일반적으로 통계조사업무는 필요성이나 작성능력이 대규모로 이루어지는 점에서 정부나 지방자치단체 차원에서 수행되는 관청통계라는 특성을 갖고 있다. 이러한 점 때문에 통계조사와 관리능력은 선진국과 개발도상국을 구분 짓는 중요한 기준이 되고 있다.

국가는 다양한 관점에서 생산된 통계 데이터를 이용자 편의에 따라 정확하게 해석하여 사회현상을 이해할 수 있도록 모든 통계 데이터를 통합관리할 의무를 갖고 있다. 또한 도서관은 이러한 통계 데이터를 이용자에게 적합한 형태로 제공해야 하는 책임을 갖고 있다. 왜냐하면 방대한 양의 통계 데이터를 인간의 기억과 수작업으로 개인적으로 처리하거나 해석하기에는 상대적으로 어렵기 때문이다. 따라서 통계 데이터도 일반 도서관 자료와 같이 대규모 데이터를 효과적으로 관리 및 검색하기 위해서 해당 정보를 요약해야 한다. 도서관에서는 요약의 방법을 표준화된 기준에 의해 수행하고, 그 기준을 메타데이터라 한다. 전통적으로 정보에 대한 요약행위는 원정보(raw data) 압축에 따라 최소한의 의미적 손실을 초래하지만, 통계 데이터의 경우에는 원정보의 손실적인 측면보다 원정보에 대한 설명적 요소로서 메타데이터를 부착해야 하는 특성을 갖고 있다.

이러한 특성 때문에 많은 국가나 국제기구에서는 통계 데이터의 효율적 관리를 위해 자국의 통계 데이터를 표준화할 수 있는 새로운

통계 메타데이터를 개발하고, 이의 국제적 표준에 대한 연구를 지속적으로 수행하고 있다.

그러나 통계 데이터에 대한 요약의 수준과 표현방법은 각 주제별 혹은 국가상황의 특성에 따라 독특한 특성을 갖고 있으며, 우리나라도 표준적인 통계 메타데이터를 모든 통계 데이터에 완전하게 적용하지 않고 있다. 예를 들면, 통계청 생산 통계 데이터와 한국은행 혹은 재정경제부에서 생산하는 통계 데이터의 내용적 유사성에도 불구하고 표현형식의 상이함으로 인해 이를 통합적으로 관리하지 못하고 있다. 또한 OECD의 회원국으로서 우리나라 통계청에서 생산하는 통계 데이터와 주요 선진국에서 생산하는 통계 데이터도 표현형식이 상이하기 때문에 해외 통계정보의 활용에도 커다란 장애요인이 되고 있다. 이러한 상이한 통계 데이터를 통합하여 하나의 데이터베이스로 관리할 수 있어, 이용자가 다양하고 편리하게 활용할 수 있는 환경을 제공하기 위해서는 표준화된 통계 메타데이터 구축이 필요하다. 특히 통계 데이터는 각국에서 공개를 기본원칙으로 하기 때문에 특정 국가의 기준을 다른 국가가 수용하는 방식보다 국제표준에 자국의 통계표현 형식을 적용하는 방식을 취하고 있기 때문에 국내의 다양한 통계 데이터를 통합적으로 표현하고, 국제표준을 준수하는 통계 메타데이터 개발이 필요하다.

본 연구에서는 이를 위해 주요 선진국가 및 국제기관의 통계 메타데이터 활용현황과 관련된 가이드라인을 분석하여 표준적인 통계 메타데이터를 조사하고자 한다. 또한 우리나라 통계청에서 생산하는 국가 주요 조사보고서의 외형적 · 내용적 정보를 분석하여 공통적인 메타데이터 요소를 도출하고자 한다. 궁극적으로 이 두 개의 조사결과를 근거로 하여 국제기준을 수용하고, 국내 통계 데이터 표현특성을 반영할 수 있는 국가 표준통계 메타데이터를 제안하고자 한다.

2. 통계 데이터의 구성

통계 데이터는 함유된 의미가 숫자로 이루어진 정보로서 해당 정보의 의미를 파악하기 위해서는 반드시 그 숫자가 어떠한 의미를 갖는지에 대한 보조설명(좌표나 필드 명 등)이 있어야 한다. 따라서 통계 데이터를 해석하기 위해서는 원정보와 통계 데이터, 통계 메타데이터가 반드시 함께 존재해야 하고, 이러한 모든 요소가 충분히 구조화될 경우에만 통계 데이터로서 의미를 갖는다.

2.1 통계 데이터의 생성

통계 데이터는 특정한 목적을 갖고 질문과 응답이라는 과정을 통해 조사대상물에 관한 것을 수치로 표현한 것이다. 또한 통계조사는 통계 데이터를 획득하기 위한 목적으로 실시되어 그 결과로부터 작성되는 것과, 여러 행정상의 기록이나 보고 등에서 작성되는 것으로 대별된다. 전자의 경우를 제1의 통계라 하며, 국세조

사를 비롯하여 센서스 등 몇 개의 대규모 전수(全數) 조사와 각 부처 등에서 각각 행정의 목적에 따라 이루어지는 여러 표본조사에 의하여 작성되는 통계이다. 후자의 경우는 제2의 통계 또는 업무통계라 하며, 출생 · 사망 · 결혼 · 이혼 신고에 바탕을 두는 인구동태 통계를 비롯하여, 각 부처의 행정상 기록 · 보고를 근거로 편성되는 통계이다. 업무통계는 직접 통계조사로 구분하지 않지만, 센서스와 같이 정기적으로 이루어지고 있다. 예를 들면, 재정경제부 · 한국은행 등에 의해 편성되는 재정 · 금융통계 등 사회나 경제에 관한 중요한 통계가 많으며, 통계체계 가운데서도 중요한 위치를 차지하고 있다. 따라서 통계조사는 주체적 관점에서 조사의 예산과 규모 등과 같은 문제로, 대부분 정부나 지방자치단체 등에 의한 관청통계로 작성되는 특징을 갖고 있다.

이러한 관청통계는 정부의 각 부서별로 생산되기 때문에, 현재는 통계 데이터의 표현과 구조화에 서로 다른 기준을 사용하고 있다. 한편, 우리나라 관청통계 데이터는 구조와 메타데이터의 생성, 표현방법의 표준화 등에 대한 파급효과가 높은 특징을 갖고 있다.

2.2 통계 데이터의 종류

통계 데이터는 통계 메타데이터에 의해 제공되는 기술사항의 주된 객체이다. 통계 데이터는 실험이나 측정 또는 설문조사를 통해서 수집된 데이터를 의미한다.

통계 데이터는 통계과정에서 두 가지 의미를 갖고 있다. 첫 번째는 조사된 원정보이며, 두 번째는 하나의 수치 데이터로 변화된 수치정보이다. 따라서 통계 데이터는 정보의 손실과정에 따라 원정보와 마이크로 정보, 매크로 정보, 통계 메타데이터로 구분할 수 있다(통계청 2004). 이 가운데 앞의 세 가지의 특성을 조사하면 다음과 같다.

1) 원정보

통계기관에서 의도하는 통계 데이터를 수집하기 위해 질문조사표나 면담에 의해 수집된 원래의 정보를 의미한다. 수집되는 정보의 형태는 문자나 혹은 수치와 같은 기호로 이루어지며, 통제가 이루어지지 않아 정보의 손상이 없는 형태의 정보를 의미한다.

2) 마이크로 데이터

마이크로 데이터는 흔히 관찰 데이터 또는 측정 데이터로 불리며, 이는 어떤 대상에 대한 일련의 특성의 집합에 대한 관찰 또는 측정의 결과로 생겨나는 것이다. 예를 들면, 개체수 조사를 비롯하여 설문조사, 실험 등을 통해 수집되는 개별적인 자료를 의미한다. 원정보 형태와 의미에 가장 가까우며 수집될 대상이 갖고 있는 정보를 기호로 표현된 수준의 데이터이다.

3) 매크로 데이터

매크로 데이터는 합계와 평균, 빈도수 등과 같이 마이크로 데이터를 다시 가공한 형태의 데이터이다. 즉, 최종적으로 가공된 데이터로서 이용자로 하여금 해당 데이터로 원정보를 추론하고, 통계 목적에 맞는 예측을 가능하게

하는 수준의 데이터이다. 매크로 데이터는 마이크로 데이터를 근거로 작성된다.

2.3 통계 메타데이터

정보의 추상화를 위한 메타데이터는 원정보를 요약한 것과 원정보를 기술하는 것으로 구분할 수 있다. 전자의 대표적인 것으로 MARC(Machine Readable Catalogue)가 있다. 이는 인쇄형 책자와 같은 문헌을 구조화하기 위해 해당 책이 갖고 있는 주요 특성[1]을 기준으로 작성한, 데이터베이스를 작성할 수 있는 구조정보이다. 이는 원정보를 요약한 것으로 원정보가 갖고 있는 정보가 손실된 형태이다. 후자의 경우는 통계 메타데이터로서 이용자가 접할 수 있는 최종수치 데이터가 갖고 있는 정보를 기술지원(descriptive assist)하는 메타데이터이다. 따라서 이 정보는 통계 데이터에 대한 포괄적인 정보를 보다 쉽게 접할 수 있는 게이트웨이를 의미한다. 이는 전자와 달리 수치화에 따라 손상된 원정보를 지원해주는 역할을 수행한다(UN 1999).

이러한 메타데이터의 특성 가운데 통계 메타데이터는 기존의 메타데이터나 기술표준에 비해 원정보 손상을 보전한다는 특성에 따라 별도의 위상을 갖는데, 이는 다음과 같은 통계관리의 특성 때문이다.

첫째, 통계적 분석과정은 수학을 데이터에 적용하여 실세계를 표현한다. 그러므로 통계적 분석에 있어서 실세계의 현상을 구분하고 인공적인 명명(命名) 또는 기호변환을 적용한 숫자코드로 분류하는 것이 관습적으로 이루어진다. 이러한 요구는 초기통계 패키지에서 응용되었던 방법처럼 해당 패키지 내에 저장되어 있는 숫자를 해석하기 위해서 몇몇 종류의 메타데이터가 필요하다는 인식에 기인한다. 전통적인 패키지는 '변수 레이블'과 '값 레이블'을 지니고 있기 때문에 해당 데이터에 의미를 부여할 수 있었다. 통계 데이터는 일반적으로 각 축에 의미를 부여한 수에 대한 직사각형 또는 정육면체처럼 여겨져 왔다.

두 번째로 통계분석은 수많은 주제영역에 적용될 수 있다. 그러나 통계정보 시스템이 이러한 개념을 통계정보 시스템의 프로세스에 포함하고 있지 않기 때문에 해당 영역의 개념을 설명할 필요가 있다.

세 번째로 통계적 처리를 활용하는 많은 주제 전문가들이 고급 통계학자는 아니므로, 데이터를 정확하게 해석하기 위해서는 통계적 개념과 복잡한 요소들에 대한 설명이 필요하다.

통계 메타데이터는 다른 메타데이터가 '데이터에 관한 데이터' 혹은 '데이터의 데이터'로 표현될 수 있으나, 통계 메타데이터는 '포장 데이터(imputed data)'와 '숨어있는 데이터(behind data)'로 표현하는 것이 보다 정확하게 설명될 수 있다. 통계 메타데이터는 통계의

1) 저자명, 서명, 책의 크기, 페이지 등

결과물로 생성된 통계 데이터가 그 자체만으로는 통계 데이터가 함축하고 있는 현실세계를 충분히 설명할 수 없기 때문에 이에 대한 추가 기술정보가 요구된다. 따라서 이러한 추가 기술정보는 원정보로부터 추출되어지는 것이 아니며, 오히려 원정보에 부가되는 성격을 띠고 있다. 그러므로 통계 메타데이터는 원정보보다 정보의 양이 많을 가능성이 있으며, 서지 데이터에 비해 정보가 보다 설명적, 기술적이라 할 수 있다.

2.3.1 통계 메타데이터의 기준

인터넷 상에서 통계자료를 사용하는 이용자는 초심자로부터 전문가에 이르기까지 매우 다양하다. 이러한 다양성을 인정하면서 이용자에게 필요한 통계 메타데이터는 세 가지 관점에서 최소한의 기준을 필요로 하고 있다(UN 2000).

1) 인터넷 검색 및 탐색을 위한 메타데이터

이용자의 검색 능력과 활용능력을 고려하여 통계정보를 제공하는 사이트에서 요구되는 최소한의 메타데이터는 다음과 같다.

사이트의 사이트 맵과 내용 테이블

FAQ

새로운 조사결과물에 대한 사이트 뉴스

통계주제 분야에 대한 설명

통계기관 설명

통계 시스템에 대한 설명

생산자료에 대한 개략적 설명

담당자의 연락처

조사결과 발표 예정표(예정 일정표)

주요 통계 사이트와의 링크

웹 상의 정보를 구매나 구독을 위한 연락처 정보

2) 통계자료 해석을 위한 메타데이터

통계관리기관에서 제공되는 정보해석에 대한 필요성은 주제영역과 정보형태, 이용자 수준에 따라서 달라질 것이다. 따라서 통계관리기관에서는 가능한 모든 이용자를 위해 중요하다고 판단되는 데이터에 대해 설명을 부기할 필요가 있다. 이러한 설명요약문은 통계자료 이해에 오류를 최소화할 수 있을 것이다.

타이틀/내용 설명 : 이 부분에 포함될 대략적인 요소는 다음과 같다.

- 조사대상 인구
- 지역정보
- 관리부서
- 표준분류체계

테이블이나 그래프의 행과 열에 대한 레이블

레이블의 정의

측정단위

피조사 기관

지역 단위

누락기간 (누락 데이터, 누락목록 등)

주석

정보원

테이블의 표준기호에 대한 설명

저작권과 관련된 정보

부가정보를 위한 연락처

3) 후처리를 위한 메타데이터

이 형태의 메타데이터는 통계 데이터를 다운로드나 저장할 경우를 고려해야 한다. 후처리를 위한 메타데이터는 파일의 형식과 포맷 등을 들 수 있다. 일반적으로 ASCII에 기반한 text 파일과 범용적인 csv 파일 등과 같은 파일 정보를 이용자에게 제공한다. 따라서 후처리를 위한 메타데이터는 다음 두 가지 조건을 충족시켜야 한다.

메타데이터는 앞에서 설명한 메타데이터를 포함하는 정보를 갖고 있거나, 별개 작업을 위해 자료에 쉽게 접근할 수 있어야 한다. 즉, 다운 받은 정보를 이용자가 자신의 필요에 맞게 해당 데이터를 손쉽게 가공할 수 있는 형태이어야 한다.

데이터/메타데이터는 다른 통계 소프트웨어를 이용하여 손쉽게 처리할 수 있는 호환성과 이식성을 가져야 한다. 이를 위해서는 특정 소프트웨어에 의존하지 않는 ASCII 코드로 이루어진 데이터 형태로 존재할 필요가 있다.

2.3.2 통계관련 데이터의 수준

통계관련 데이터는 수준에 따라 다음 네 가지 종류로 통계 메타데이터의 수준을 구분할 수 있다.

통계 데이터 : 통계자료의 공유를 비롯하여 검색, 이해의 수준으로 구조화된 데이터를 의미한다.

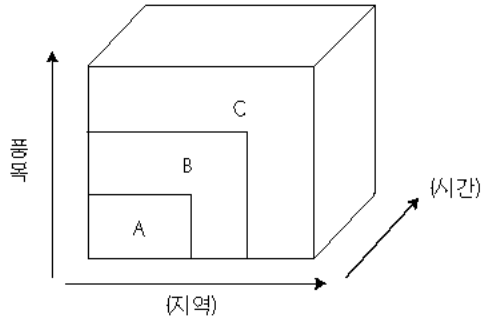
시스템 메타데이터 : 수집된 자료에 대한 물리적 정보로서 데이터베이스를 구축하는 수준의 데이터를 의미한다. 이에 속하는 데이터로는 자료의 저장위치를 비롯하여 레코드의 형태, 데이터베이스의 구조, 매체·자료의 크기 등과 같은 시스템적인 수준이다.

응용 메타데이터 : 통계 자료의 내용 및 사용과정에 대한 서술적 설명을 지원하는 수준의 데이터이다. 이에 속하는 데이터로는 질문 조사표의 형식을 비롯하여 사용되는 변수에 대한 정의, 통계 데이터를 통제하기 위해 사용되는 소프트웨어에 관한 정보, 자료편집에 관한 설명 등을 지원하는 수준의 데이터이다.

관리 메타데이터 : 구축된 통계자료를 관리하는 과정에서 생산된 정보를 의미한다. 이에 속하는 정보로는 해당 통계조사를 실시하는 데 소요된 예산을 비롯하여, 예산의 소항목으로 사용된 경비, 조사일정, 조사결과를 분석하는 데 소요된 일정, 조사된 자료의 유효기간 등과 같이 조사과정에서 생산된 관리용 데이터를 의미한다(권용수 1998).

2.3.3 통계 데이터 간의 연관성

통계 데이터는 최종적으로 제시되는 형태가 수치형이기 때문에 최종적으로 표현되는 통계수치는 당연히 손상된 정보를 갖고 있다. 즉, 별도의 설명이 없을 경우 해당 통계 데이터는 숫자로 이루어진 기호일 뿐이다. 이 수치 데이터는 완전하게 요약된 정보로서 통계 메타데이터와 같은 속성(범주) 정보를 부착하고 있어야 한다(박원환, 황조연 2004). 이 관계를 예를 들어 설명하면 다음과 같다(<그림 1> 참조).



〈그림 1〉 통계 데이터의 구조

〈표 1〉 일반 및 통계 메타데이터 특성 비교표

특 성	일반 메타데이터	통계 메타데이터
정보원	원정보	가공정도가 높음
데이터의 유형	문자 중심	숫자 및 문자 중심
원정보와의 관계	독립적	의존 및 종속적 특성
정보의 속성	범주속성 중심	범주 및 요약속성 중심

예) 〈2004년 8월 1일 대전청사에서 서울까지 우등고속 버스 요금은 10,000원〉

이 예에서 얻을 수 있는 범주속성 정보는 시간(2004년 8월 1일)과 지역(대전청사 서울), 우등고속(품목)이 해당한다. 그리고 실제 얻어지는 버스 요금인 10,000원은 요약정보에 해당한다. 우등고속 버스 요금인 10,000원은 수치 데이터만으로는 어떤 의미인지를 알 수 없기 때문에 반드시 범주속성 정보를 갖고 있어야 한다. 이때 10,000원은 원정보 혹은 통계 데이터이며, 범주속성 정보는 메타 정보에 해당한다. 이를 통계 데이터로 표현하면 10,000원(C)은 단순히 운임을 표현하는 수치로 구성된 최종 조사된 집계/요약자료이다. 이 정보는 수치정보만으로 표현되는 특징을 갖고 있으며, 단독으로는 의미를 나타내지 못하는 자료이다. B는 가공처리 데이터(imputed data)로서 가공되지 않은 조사자료를 범주속성(버스 출발일자, 지역정보, 품목정보 등) 정보와 요약자료 등이 연결된 정보이다. A는 해당 시간(일자)에 대전발 서울행(지역 혹은 공간) 우등고속 버스 요금(앞의 예)을 표현한 원정보이다(통계청 2004).

따라서 이상과 같이 통계 메타데이터를 조사 · 분석한 결과 일반 메타데이터와 통계 메타데이터를 비교분석하면 〈표 1〉과 같이 구분할 수 있다.

일반 메타데이터는 원정보에 대해 매우 독립적이어서, 원정보와 별도로 해당 데이터만으로도 이용자에게 최소정보를 제공할 수 있으나, 통계 메타데이터는 원정보에 대한 의존 정도가 상대적으로 매우 높다. 예를 들면, MARC에서 저자명은 그 자체만으로 정보를 함유하고 있으나, 통계 데이터에서는 통계 데이터와 통계 메타데이터는 독립적으로 이용자에게 정보를 제공하지 못하고 있다. 즉, 통계정보의 경우, 해당 정보에 대한 메타 정보를 이용자에게 제공하지 않을 경우 정보의 가치가 급격히 낮아진다. 따라서 통계 데이터가 정보적 가치를 갖기 위해서는 반드시 통계 메타데이터가 존재해야 한다(Colledge, Wensing, and Brinkley 1996).

2.4 통계정보 시스템

현대 정보화 사회에 있어 통계 데이터를 일관되고 체계적으로 수집하고, 이를 통합적으로 관리할 수 있는 통계정보 시스템을 국가차원에서 개발 · 운영하고 있다. 각 국의 통계정보 시스템은 기능상 차이는 있으나 기본적인 기능과 시스템 구성은 대동소이하다.

2.4.1 기 능

통계정보 시스템은 입력 수집과정을 비롯하여, 축적 처리과정, 결과 배포과정과 같이 3단계 과정을 거쳐 통계 데이터를 처리한다.

1) 입력 수집과정

'입력 수집과정'은 통계 데이터를 수집하기 위한 기획 및 데이터 수집준비, 관측 등록 등과 같은 세부과정으로 구성되며, 각 단계별 세부 처리내용은 다음과 같다.

조사준비

데이터 수집

데이터 준비

관측 등록 완료

2) 축적 처리과정

'축적 처리과정'에는 입력된 수집정보를 축적하여, 통계 모델을 생성하는 모델링 과정과 해당 결과를 추정하는 추정과정을 포함한다.

3) 결과 배포과정

'결과 배포과정'은 축적 처리과정을 거쳐 생성된 결과물을 대외로 공표하는 과정이 포함된다. 전통적으로 이 과정은 인쇄물에 기반한 출력물에 의존하였으나, 최근에 인터넷과 같은 가상공간에 공표하는 것이 세계적인 추세가 되고 있다. 이 과정의 세부과정으로는 어떤 형식으로 처리된 결과를 표현할 것인지를 결정하는 공표과정과 이를 대외적으로 배포하는 과정이 포함된다.

2.4.2 구 성

통계정보 시스템을 구성하는 것은 일반적인 시스템 구성과 같이 통계정보 및 통계정보 시스템을 필요로 하는 이용자와 통계정보를 생산하는 생산자, 해당 통계정보를 서비스하기 위한 메타데이터로 구성된다.

1) 이용자

〈표 2〉 통계정보 시스템의 이용자 유형

이 용 자	이 유
정 부	어떤 행동에 대한 계획, 감시, 평가
회사, 조직	경영적인 의사결정
정 치 인	협상, 로비
연 구 자	현실세계의 현상에 대한 분석, 이해, 설명
국 민	민주적 과정에 참여

통계정보 시스템을 이용하는 사용자는 사용 목적에 따라 크게 세 개의 군으로 구별할 수 있다. 첫 번째 이용자군은 특정한 의문과 문제를 갖고 있는 잠정적인 이용자로서, 자신이 갖고 있는 어떤 문제를 해결하기에 적합한 통계정보를 찾는 이용자군이다. 두 번째 이용자군은 자신의 흥미와 관련된 어떤 통계정보를 찾고, 이러한 데이터를 검색하고자 하는 이용자군이다. 세 번째 이용자군은 통계정보를 분석하고 해석하기 위한 이용자들로서 이들은 반복적으로 탐색과 검색, 분석의 과정을 수행하는 이용자군이다. 첫 번째 이용자군은 통계정보 활용숙련도 측면에서 통계 데이터 분석과, 활용적 측면에서 초심자에 속한다. 두 번째 이용자군은 통계에 대한 일반적인 지식과 이를 기반으로 특정한 내용을 입증하거나, 부연 설명을 위해 사용하는 이용자이다. 세 번째 이용자군은 전문 통계학자로서 통계정보를 자신의 의도에 따라 자유롭게 정보를 가공할 수 있는 수준의 이용자이다.

위와 같이 이용자들을 대상으로 통계정보 시스템에서 이용자들에게 적합한 통계정보를 제공하기 위해서는 각 이용자 수준에 맞는 적절한 메타데이터가 필요하다. 이러한 이용자들의 예는 〈표 2〉와 같다(통계청 2004).

2) 생산자

협의의 통계정보의 생산자는 통계정보 시스템의 입력 수집 단계에 관여하는 생산자를 의미한다. 광의의 통계정보 생산자는 통계정보의 생산 전과정에 관여하는 생산자를 의미한다. 일반적으로 통계정보 생산자를 지칭하는 것은 후자에 속한다. 생산자의 일반유형을 조사하면 다음과 같다.

통계조사와 통계정보 시스템의 설계자 : 주제분야 통계학자, 통계 방법론자, 정보 시스템 전문가 등

입력 정보 제공자 : 응답자, 접촉한 사람, 조사원 등

생산 통계학자 : 입력 수집 단계의 생산자

3) 통계 메타데이터

대부분의 통계정보 생산자들은 각각의 수준에 맞는 메타데이터를 필요로 한다. 예를 들면, 생산통계 학자들은 통계의 생산단계가 얼마나

적절하게 잘 수행되고 있는지를 확인하고, 새로운 직원을 교육하기 위해서 여러 체크 리스트와 생산 시스템 문서를 필요로 한다. 통계정보 시스템을 평가하는 감독은 그 시스템의 기능과 관련한 메타데이터를 비롯하여, 이용자들로부터 피드백 정보를 포함하는 메타데이터 등을 필요로 할 것이다.

이처럼, 통계 메타데이터는 광의의 통계정보의 생산에 참여하는 모든 이용자들에게 필요한 정보를 제공하기 위한 하나의 도구로서 활용될 수 있다(통계청 2004). 위와 같은 기능을 수행하기 위해 통계정보 시스템은 다음과 같은 주요 요소들을 기반으로 구축된다(이재창, 전명식, 김은석 1997).

- 통계자료 데이터베이스
- 메타데이터 저장소(metadata repository)
- 조사에 관련된 문서 저장소(survey document library)
- 사용자 접근 부문(user interface)
- 자료 조작 및 분석 도구모음
- 메타데이터/문서의 제작 및 갱신도구
- 다른 정보 시스템으로의 접근통로

3. 통계관련 메타데이터 요소분석

3.1 우리나라 통계청의 메타데이터 요소

우리나라 통계청은 국가차원의 현상과 수준을 분석하기 위해 조사통계 42종을 비롯하여, 가공통계 10종, 보고통계 1종을 포함하여 총 53종의 통계작성 조사업무를 수행하고 있다. 이 조사보고서는 10년마다 이루어지는 '국부통계조사'를 비롯하여, 5년 주기로 이루어지는 '인구총조사' 등 연간, 분기, 월간으로 국가현황을 통계수치로 발간하고 있다. 발간된 자료는 크게 인쇄형과 온라인형, 인쇄 및 온라인 동시 발간형으로 구분하고 있으며, 최근에는 대부분의 통계조사 결과물이 온라인 형태로 발간되고 있다. 특히, 1991년부터 KOSIS(KOrean Statistical Information System) 통계정보 시스템을 통해 통계청 생산 통계자료를 포함하여 국내 및 UN, IMF 등의 주요 통계정보를 서비스하고 있다. 본 장에서는 온라인과 책자형으로 발간되는 주요 조사보고서를 분석하여 메타데이터의 표준요소를 도출하고자 한다. 분석대상은 최근 3년간 발행된 53종의 조사보고서 가운데 구조 및 형태적 요소의 중첩성을 최소화하여 총 32종(112권)을 선정하였다.

통계청에서 생산되고 있는 통계자료집은 크게 조사표와 수주통계 데이터, 인구조사로 구분할 수 있다. 이상의 자료는 형태적으로 표제를 비롯하여 머리말, 이용자를 위하여(일러두기), 차례(목차), 조사개요, 조사결과 요약(개요), 통계표, 부록과 같은 공통적인 메타요소를 가지며, 참조정보와 같은 독립적인 메타요소를 갖고 있다.

3.1.1 공통요소

통계청 조사보고서는 8개의 유항목과 23개의 강항목으로 구성되어 있으며, 개략적인 정보

〈표 3〉 통계청 주요 조사보고서의 메타 항목

제1수준(유항목)	제2수준(강항목)
표제지	등록번호, 연도, 제목, 항목, 조사범위, 발행기관, 발행일
머리말	
이용자를 위하여	부호설명, 연락처
차례(목차)	
조사개요	조사목적, 연혁, 조사의 법적 근거, 조사 기준시점 및 조사기간, 조사체계, 조사범위 및 대상, 조사항목, 조사방법, 조사수행 조직, 집계 및 공표, 용어해설, 표본설계
조사결과 요약(개요)	
통계표	
부록	조사표, 통계청 발간 간행물 목록

는 〈표 3〉과 같다.

부록은 인쇄형 조사보고서에서만 확인할 수 있는 정보로서, 조사표 용지와 통계청발간 간행물 목록으로 구성되어 전자형태의 보고서에서는 간행물 목록정보는 제외될 수 있다.

3.1.2 독립요소

조사보고서의 메타 요소 가운데 형태적으로 상이한 것을 조사하면 다음과 같은 특정 조사보고서에 독립요소들이 독자적으로 조사되었다.

'도 · 소매업 판매액 지수' → '도표', '참고'

'한국직업표준분류' → '고시', '총설', '분류항목표'

'2003 물가연보' → '소비자물가지수 개요', '2003년간 소비자 물가'

'도 · 소매업 통계조사 보고서', '서비스업 통계조사보고서' → '표본설계 개요'

'기계수주 통계' → '설비투자 추계지수 기계류 내 수출하기 기계류수입'

3.1.3 통계청 내부 확장 메타데이터 기준

통계청에서는 이상의 결과에 근거하여, 내부적인 통계보고서 작성과정의 산출된 정보 가운데 메타데이터로서 의미를 갖는 요소를 부가하여 자체적인 통계 메타데이터 확장요소를 개발하였다(통계청 2004). 통계청 확장 메타데이터는 크게 조사개요와 조사방법론, 자료제공, 조사표, 용어해설, 관련문서정보, 조사내용 품질, 예산 및 인력관련 요소, 기타 등 9가지 대항목과 32개의 강항목, 49개의 목항목, 4개의 세목항목 등 총 85개의 항목으로 구성되어 있다. 각 항목은 1수준(유항목) 4수준(세목)으로 구성되어 있다.

① 조사개요 : 11개 강항목, 10개의 목항목

② 조사 방법론 : 9개 강항목, 39개의 목항목, 4개의 세목항목
③ 자료 제공 : 4개 강항목
④ 조사표
⑤ 용어 해설
⑥ 관련문서 : 조사표의 PDF 등 전자 문서
⑦ 품질 : 2개 강항목
⑧ 예산 · 인력 · 교육 · 홍보 : 4개 강항목
⑨ 기타 : 2개 강항목

이 기준과 앞의 통계청 조사보고서 항목과 중첩도를 분석하면, 앞의 항목 대부분이 이 기준에 포함되나, 머리말과 조사결과 요약(개요), 부록은 포함되지 않는다. 머리말과 부록은 책자형 자료의 고유한 특성으로서, 통계정보로서 중요도가 상대적으로 떨어져서 메타 항목으로 선정을 유보할 수 있다. 한편, 조사결과 요약은 해당 자료의 개략적인 자료로서 신문기사 자료 등과 같은 용도로 사용될 수 있어 메타 항목으로 주요한 역할을 수행할 수 있다.

3.2 SDMX와 SDDS

SDMX(Statistical Data and Metadata eXchange)는 세계은행(World Bank)을 비롯하여 BIS, ECB, IMF, EUROSTAT, OECD, UN 등과 같이 세계 주요 국제기구들이 생산하는 통계 데이터와 메타데이터를 상호 교환하기 위해 2002년부터 공동으로 개발한 통계정보 메타데이터의 국제표준안이다(ECE 2004).

이 표준안은 세계 국가와 국제기구에서 생산되는 모든 데이터를 효과적으로 상호 공유하기 위해 설계되었다. 설계의 목적은 크게 두 가지로 구분한다. 하나는 각 국에서 생산된 정보를 효율적으로 관리할 수 있는 방안이고, 나머지는 체계적으로 관리된 정보를 효과적으로 공유할 수 있는 메타데이터 표준안을 개발하는 것이다. SDMX도 후자의 관점에서 개발한 메타데이터 표준용어(MCV: Metadata Common Vocabulary)를 갖고 있다. MCV는 국가들과 국제기구에서 생산하는 통계 데이터에 대한 포괄적인 메타데이터에 대한 표준용어집으로서, 각 국이 생산하는 통계 데이터를 상호간에 효율적으로 이해할 수 있도록 설계되어 있다. MCV의 목표는 다양한 통계 데이터의 용어를 통일하여 통계 데이터 교환을 위한 메타데이터의 표준 템플릿을 제공하는 것이다. 즉 메타데이터 구성에 동일한 용어를 사용함으로써 상이한 통계정보 시스템 간 통계 데이터교환을 간단하고 용이하게 처리할 수 있도록 유도하는 것이다. 한편 IMF는 세계 각국에서 생산하고 있는 주요 전문 데이터에 대한 데이터 교환을 위한 표준안(SDDS: Special Data Dissemination Standard)을 개발하여 국가간 데이터 교환의 표준으로 활용할 수 있도록 하였다. 특히, SDDS 내에는 통계 데이터 교환을 위한 스키마를 갖고 있으며, MCV의 항목과 완전한 대치가 가능하다. 따라서 SDDS의 구조에 MCV 용어를 매칭하여 국제통계 메타데이터 표준안을 설계할 수 있다.

실제 가능성을 확인하기 위해 MCV의 용어

중에서 SDDS와 직접 관련된 메타데이터를 대치한 결과, 세부항목 요소는 6개의 유항목(제1수준)과 22개의 강항목(제2수준), 60개의 목항목(제3수준), 78개의 세목항목(제4수준)으로 구성되어 있다. 세목항목은 대부분 모든 통계조사에서 생성된 정보를 위한 것보다 각 국의 다양한 통계 데이터를 반영할 수 있도록 고려한 것이다. 또한 전개수준은 4개의 수준으로 구분되어 있으며, 이 가운데 1, 2수준 전부와 3수준 일부를 조사하면 〈표 4〉와 같다(Pellegrino and Ward 2004).

3.3 메타데이터 선정의 제한점

앞에서 통계청 발간자료와 통계청 메타데이터 요소, 국제표준 통계 메타데이터 요소 내용을 조사한 결과, 다음과 같은 몇 가지 제한점을 도출할 수 있었다.

첫째, 공통적인 메타 요소로서 서지 데이터에 속하는 표제와 목차와 같은 요소들과 통계자료에 대한 설명과 안내 관련 요소는 모든 분석 대상에서 확인할 수 있었다. 단, 이런 요소를 표현하는 용어는 자료의 시기나 목적과 형태에 따라 같은 의미를 갖고 있는 부분이라 하더라도 용어가 다르므로 이런 용어들의 통일이 반드시 필요하다. 이 작업이 선행되지 않으면 이형동의어에 속하는 여러 가지의 메타데이터 요소가 선정될 수 있는 위험이 있을 뿐만 아니라 궁극적으로 자료해석에 오류를 야기할 수 있다.

둘째, 위의 조사결과, 우리나라 통계청 자료에서만 독특하게 출현하는 독립요소를 메타데이터 요소로서 선정할 경우에 자료의 분별력과 검색은 정확하게 이루어질 수 있지만, 그 요소를 모두 메타데이터의 요소로 선정하기에는 그 양이 방대하고, 모든 관련자료의 내용을 전부 조사하는 것은 현실적으로 어렵다. 그러므로 이렇게 하나의 자료의 종류에 있거나, 적은 양의 자료에 존재하는 항목들을 메타데이터의 요소로 선정하기에는 비경제적이고, 더욱이 메타데이터의 표준요소로 선정할 수 없다.

셋째, 자료의 내용적인 면의 메타데이터 요소를 추출하기 위해, 통계자료의 내용에 해당하는 자료의 원천인 조사표를 분석해본 결과, 조사표는 자료에 따라 하나의 조사표만 있는 것이 아니라, 한 자료에도 여러 개의 조사표가 있음을 알 수 있다. 따라서 이런 조사표들에 해당하는 모든 요소를 공통적으로 추출하는 데는 문제가 있다. 그리고 조사표의 질문들을 메타데이터의 요소로 하기에는 그 형태나 내용에 있어서 비정형적이고, 문장의 형태를 갖고 있는 경우도 많기 때문에, 내용적인 면에서 공통적이고 세부적인 메타데이터를 추출하는 것도 현실적으로 어렵다.

넷째, 일반적으로 공통적으로 여러 종류의 자료에 들어있는 요소들을 메타데이터의 요소로 선정하지만, 이럴 경우, 소수의 자료에 들어있는 요소이지만 이용자의 편의를 위해 반드시 필요하거나, 많은 도움이 될 수 있는 요소들을 놓칠 수도 있다. 따라서, 단순히 각 자료에 출현한 빈도수만으로 공통의 메타데이터를 요소 선

〈표 4〉 MCV와 SDDS의 매핑 통계 메타데이터 표준안

제1수준 1	제2수준	제3수준
연락처 (CONTACT)	담당자와 부서명	성
		명
		직 위
		조직명(Organization)
		부 서
		주 소
		전화번호
		팩스 번호
		이메일 주소
데이터 (DATA)	국가-지역	국 가 명
	데이터 종류	
	데이터 특성	정보원 형태, 조사기간 외
	발행주기(Periodicity)	
	실제 발행일자(Timeliness)	
공표방법 (ACCESS by The Public)	발표 일정표	내부 일정
	유관기관 동시발표	발표 방법
	공표 데이터 포맷	출판 목록
		일/주/월/분기/연간 고시
		디 스 켓
		CD-ROM
데이터 무결성 (INTEGRITY)	책임 감독관청	해당 법률 - 통계자료
		해당 법률 - 통계신뢰도
	정부(비공개 정보) 인증	기관 (Agency)
	공개 통계 자료의 정부 인증	기 관
	데이터 공개상태, 공개자료 수정에 따른 법률적 규정	데이터 공개수준
		개정 정책
		조사방법 변경 공시
데이터 품질 참조정보 (Quality References)	정보원 추출을 위한 도구(방법론) 및 데이터 출처	표제와 출판 빈도
	참고 및 관련자료	품질보증, 데이터 형태 외

(다음 페이지에 계속)

방법론 (METHODOLOGY)	개념과 정의	국제적 지침
	데이터 취급 범위	지리적인 범위
		영역범위
		처리범위
	분 류	국제적 지침
	처리 기록	회계와 예외 조항
		평 가
		총합 절차 (Grossing procedures)
	기초 통계 데이터의 특성	데이터 수집
		출처 데이터의 평가
		항목/생산/보고단위
	통계 데이터 편집 및 가공 (Compilation practices)	조정, 추정, 가중치 외
	기타 요소	기타 주요 요소

정할 것이 아니라, 모든 요소들을 고려하여 이용자에게 필요하거나 그 자료를 설명할 수 있는 요소들을 또한 선정해야 한다.

4. 국가표준 통계정보 메타데이터 개발/국가 통계표준 메타데이터 설계

우리나라는 UN을 비롯하여 OECD에 가입하고 있어 회원국으로서 맡은 책임을 수행해야 한다. 그 가운데 국가 통계의 공표와 주요 국가 및 국제기구와의 통계정보의 활용은 미래 선진국으로 진입하기 위한 필수적인 조치이다. 이는 우리나라 통계 데이터와 메타데이터의 공표와 기술(description) 수준이 해외의 표준을 수용해야 함을 의미한다. 주요 선진국도 자국의 통계 데이터베이스 구축을 국제표준에 따르도록 노력하고 있다. 본 장에서는 앞 장에서 조사된 결과에 따라 국가 통계표준 메타데이터 요소를 제안한다.

4.1 주요 국가의 통계관련 메타데이터

4.1.1 미 국

미국 통계청은 1995년부터 통계 메타데이터와 조사표 표준화를 위한 연구와 투자를 지속하여, 조사 및 통계용 메타데이터 표준(SDSM: The Survey Design and Statistical Methodology Metadata Standard)을 개발하

였다. 이 표준에서는 이용자가 통계정보를 해석하기 위해 필요한 메타데이터 분류의 방법과 메타데이터의 의미분류 방법을 제공한다. 또한, SDSM은 조사 혹은 센서스의 데이터 보급, 분석, 처리, 설계에 대한 자료를 포함하여 모든 개념의 개요, 용어의 계층적인 시소러스도 개발함으로써 웹 정보검색의 효율성까지도 고려하였다.

SDSM의 메타데이터 항목들은 제1수준인 장(chapters)과 제2수준인 영역(sections)으로 구성된다. 각 장과 영역은 다음과 같은 메타데이터의 논리적인 구조를 갖고 있다.

다음은 SDSM이 갖고 있는 메타데이터의 장과 해당 정의를 요약한 것이다(Gregory etc. 1996).

① 확인(Identification)(필수 요소) 필수적인 메타데이터 항목의 최소의 세트를 포함.

② 내용(Content)(선택 요소) 조사의 주제 데이터의 특성에 관한 자료 포함.

③ 계획(Planning)(선택 요소) 조사 프로젝트 계획과 관련된 자료 포함

④ 디자인(Design)(선택 요소) 조사표 설계와 처리에 관련된 자료 포함

⑤ 실행(Implementation)(선택 요소) 조사과정 및 처리에서 기록을 유지하고 만들기 위한 수단 등과 관련된 자료 포함.

⑥ 분석(Analysis)(선택 요소) 모든 통계처리와 관련된 자료 포함.

⑦ 데이터 처리(Data_Processing)(선택 요소) 컴퓨터 처리에 필요한 자료 포함.

⑧ 데이터(Data)(선택 요소) 데이터와 관련된 자료 포함.

4.1.2 노르웨이

노르웨이는 아래와 같이 통계정보를 제 1수준의 메타데이터로서 관리정보를 위한 메타데이터, 조사배경과 목적을 위한 메타데이터, 통계 생산물을 위한 메타데이터, 개요와 변수, 분류를 위한 메타데이터, 에러와 불확실성 유형을 위한 메타데이터, 유사성과 일관성을 위한 메타데이터, 접근성을 위한 메타데이터 등 7개의 유수준으로 구분하고 있다. 각각의 항목은 제 2수준의 메타데이터 요소를 갖고 있다. 예를 들면, 관리정보를 위한 메타데이터로는 이름, 빈도수, 지역 레벨, 주제 그룹, 구분, 권한, UN 규정, 국제 보고 항목 등을 사용하고 있다. 다음은 노르웨이 통계청에서 채택하고 있는 통계 메타데이터의 제1수준이다(Hustoft, Linnerud, and Sebo 2004).

① 관리정보

② 조사배경 및 목적

③ 통계 생산물

④ 개요 및 변수, 분류

⑤ 에러 및 불확실 유형

⑥ 유사성 및 일관성(Comparability and coherence)

⑦ 접근성

4.1.3 캐 나 다

캐나다는 1998년부터 캐나다 정부에서 조사

및 생산하는 약 400여개의 조사에서 IMDB(Integrated Meta Database)를 구축하여 대국민 정보제공 서비스를 실시하고 있다. 캐나다에 있어서 메타데이터 활용능력을 극대화하여 민간기업의 경쟁력과 함께 개인 이용자의 정보 활용도를 최대화하기 위해 노력하고 있다. 메타데이터는 다음 세 가지 기본기능을 유지하기 위해서는 다음 세 가지 역할을 수행해야 한다. 데이터 배포, 데이터 생산, 통계 시스템의 관리, IMDB는 기본적으로 이용자가 해당 데이터를 이용할 때 원자료의 접근과 해석을 지원하기 위한 것이다. 다음은 캐나다의 메타데이터 표준요소 수준 가운데 제1수준을 정리한 것이다. 7개의 유항목과 24개의 강항목, 50개의 목항목으로 구성되어 있으며, 거의 모든 요소가 SDDS의 메타요소로 표현이 가능하다. 또한 UN에서 제안하는 메타데이터 최소수준(UN 2000)을 대부분 만족시키고 있다. 다음은 캐나다의 통계정보 시스템에서 확인할 수 있는 메타데이터 요소 가운데 유항목 수준만을 열거한 것이다.

① 조사 개요
② 조사의 특성(survey characteristics)
③ 조사 시기(survey cycle time frames)
④ 방법론 (Methodology)
⑤ 조사표 (Questionnaires)
⑥ 키워드 (Keywords)
⑦ 도큐멘테이션 (Documentation)

4.2 통계정보 표준 메타데이터 개발

주요 국가에서 제안하는 표준 메타데이터는 수준과 해당 항목의 용어선택에서 국제표준과 많은 차이를 보이고 있다. 세부적인 하위 항목에서도 국제표준과 의미적으로 대부분 일치하고 있으나 전체적인 구조는 많은 차이를 보이고 있다.

한편 국내 통계정보도 국제표준을 기준으로 수준과 용어선택에 많은 차이를 보이고 있다. 이러한 점을 극복하기 위해서는 국제기준을 수용하고, 우리 고유의 통계항목을 수용하기 위한 표준 메타데이터는 더블린 코어와 같은 한정어를 이용하여 메타데이터 표준안을 설정하였다. 표준안 설정을 위해 본 연구에서는 다음과 같은 과정을 거쳤다.

첫째, 통계청에서 제공하는 주요 통계자료의 32종을 분석하여 공통의 메타데이터 요소를 추출하였다.

둘째, 이 요소를 메타데이터의 표준안 가운데 타 정의 요소와 비교하여 공통의 요소를 설정하였다. 타 정의 요소는 요소와 한정어의 두 단계로 정의되어 있는데, 타 정의 요소는 통계청 자료의 공통요소 가운데 주항목과, 한정어는 하위 항목과 같은 단계로 설정하였다. 설정된 타 정의 요소를 통계정보 메타데이터 요소의 국제기준인 SDMX의 MCV와 SDDS를 매핑한 메타데이터 요소를 통계청 내부 메타데이터 요소와 비교하여 최종적인 메타데이터 표준안을 확장하였다.

〈표 5〉 통계청 생산 주요 조사보고서의 공통 항목

제1수준(주 항목)	제2수준(하위 항목)
표 제 지	등록번호, 연도, 제목, 발행기관, 발행일
머 리 말	
이용자를 위하여	부호설명, 연락처
차 례	
조사개요	조사목적, 연혁, 조사기간, 조사대상, 조사방법, 조사항목
부 록	조사표, 통계청발간 간행물 목록

4.2.1 확장 사전단계

앞 장에서 우리나라 통계청에서 제공하는 주요 통계자료 가운데 32종을 선정하여 각 조사보고서에 표현된 모든 메타데이터 요소를 분석하였다. 제1수준에 나타난 표제를 비롯하여 6개 항목은 대부분 모든 조사대상 자료에서 확인할 수 있었다. 제 2수준은 전체 23개 항목 가운데 중복출현 빈도가 50%를 넘는 항목만을 선정한 결과 15개 항목을 선정할 수 있었다. 이상의 결과를 요약하면 〈표 5〉와 같다.

주항목과 하위 항목에서 용어간의 차이(이형동의어)는 동일한 개념으로 간주하였으며, 용어 선택은 명사형으로 통일하였다.

4.2.1 메타데이터 표준안 확장단계

이 단계는 공통요소 항목을 보편적인 더블린 코어 요소와 한정어로 표현할 수 있는 지를 확인하여 확장의 타당성을 검증하는 과정이다. 일차적으로 더블린 코어를 활용하는 것은 더블린 코어가 네트워크 자원의 검색에 적합하며, 작성이 간단한 핵심 데이터 요소를 정의하는 것으로서 레코드 작성이 용이하고, 여러 분야에서 공통으로 사용하는 핵심 데이터 요소를 규정하기 위한 용도로 사용되는 범용적인 메타데이터 세트이기 때문이다. 유럽과 미국에서는 정부차원에서 메타데이터 표준으로 더블린 코어를 채택하고, 분야별로 필요한 요소를 확장하는 방식으로 추진하고 있다.

이에 따라 통계청 조사보고서 공통 항목 가운데 더블린 코어의 15개 요소와 매핑한 결과 제목, 발행기관, 등록번호 등 3가지 항목만 일치하고, 나머지 항목을 표현하기 위해서는 한정어를 이용해야 함을 확인할 수 있었다. 통계분야에서도 메타데이터 표준으로 더블린 코어를 적용할 수 있도록 확장할 경우에는 앞서 제안한 요소들을 〈표 6〉과 같이 5개 요소는 표현할 수 있으나, 부록은 대응 한정어[2] 적용에 한계를 보이고 있다. 단, 이 가운데 '부록'과 같이 전통적

2) 김태수 안 참조.

〈표 6〉 더블린 코어에 기반한 통계 메타데이터 표준(제1수준)

기본 요소	설 명
DC.Title	표 제
DC.Description.tableOfContents *	차례(목차)
DC.Description.introduction *	머 리 말
DC.Description.note *	이용자를 위하여
DC.Description.purpose *	조사목적(조사개요)
Statistical.appendix	부 록

인 책자형 데이터에 출현한 요소는 웹 기반의 국가 통계표준 메타데이터 설계과정에는 생략할 수 있다. 음영부분(*표시)은 더블린 코어의 한정어를 이용하여 확장한 요소이다.

이상과 같이 더블린 코어의 한정어를 적용하여도 국내 통계정보의 메타데이터 표준을 제한적으로 설계하는 것이 완전하지 않기 때문에 별도의 메타데이터 자료가 필요하다. 이에 따라 본 연구에서는 SDMX와 SDDS에서 중복되는 요소를 수용하였다. 두 개의 기준을 적용함에 있어 더블린 코어의 확장 한정어와 SDMX와 SDDS의 중복 요소가 중첩될 경우, 통계정보의 메타데이터 표준인 SDMX와 SDDS에서 중복된 요소를 중심으로 기재하였다. 이는 국가 통계정보는 국제적 공유와 활용을 우선적으로 고려해야하기 때문이다. 예를 들면, 통계청의 통계 서비스 가운데에는 UN과 OECD 생성 통계정보를 통계청 홈페이지에서 동시에 제공하고 있어 이미 통계정보의 국제적 공유가 이루어지고 있다. 즉, 국내 통계정보와 해외 통계정보와의 통합부분을 고려하면 SDMX와 SDDS의 중복 요소가 다른 메타데이터보다 효율적이다.

확장결과는 43개의 요소 중 SDMX와 SDDS의 중복요소는 29개의 항목을 채택하였으며, 더블린 코어는 14개의 요소를 채택하였다. 선정된 결과는 UN에서 요구하는 통계 메타데이터 구성의 최소요구 기준(UN 1999)을 모두 수용하도록 설계하였다. 〈표 7〉은 국가 통계표준 메타데이터 요소 표준안이다.[3]

이 기준안에는 앞의 통계청 조사보고서의 기본요소와 통계청 내부용 요소 등에서 중시되는 요소를 최대한 수용하고, 국제 통계정보 공유 시스템에 참여하기 위한 요소를 더블린 코어와 SDMX와 SDDS의 중복요소를 이용하여 매핑하였다. 예를 들면, 더블린 코어의 메타 요소에서 '조사범위 및 대상' 항목은 더블린 코어와 SDMX와 SDDS의 중복요소로 모두 표현할 수 있었으나, SDMX와 SDDS의 중복요소를 채택

3) 출처 : SDM의 MCV, 〈http://www.sdmx.org/Data/SDMX_MCV_release1_200404.pdf〉

〈표 7〉 국가 통계표준 메타데이터 표준안

표준 요소	설 명
DC.Description.tableOfContents	목 차
DC.Description.introduction	머 리 말
DC.Description.note	이용자를 위하여
DC.Description.purpose	조사목적
DC.Description.procedures	조사방법
DC.Description.items	조사항목
DC.Description.results	조사결과 요약
DC.Date.surveyPeriod	조사실시 기간
DC.Date.dateOfSurvey	조사기준 시점
DC.Date.historyOfSurvey	연 혁
DC.Identifier	등록번호
DC.Relation.requires	조사의 법적근거
DC.Relation.references	용어해설
DC.Source.questionnaire	조 사 표
SDMX.Contact	발행기관
SDMX.Contact.organization	발행 부서명 또는 조사수행부서
SDMX.Contact.email	연락처에 표기된 이메일 주소
SDMX.Contact.postal	발행기관 우편주소
SDMX.Contact.telephone	연락처에 표기된 전화번호
SDMX.Contact.fax	연락처에 표기된 팩스 번호
SDMX.Data.country	국가 및 지역
SDMX.Data.category	데이터 종류
SDMX.Data.characteristics	데이터 특성
SDMX.Data.periodicity	발행주기
SDMX.Data.timeliness	실제 발행일자
SDMX.Access.releaseCalender	발표 일정표
SDMX.Access.simultaneousRelease	유관기관 동시발표
SDMX.Acess.disseminationFormat	공표 데이터 포맷
SDMX.Integrity.confidentiality	해당 법률 통계 신뢰도
SDMX.Integrity.internalAccess	정부(비공개 정보) 인증
SDMX.Integrity.ministerialCommentary	공개 통계자료의 정부 인증
SDMX.Integrity.statusOfDataUponRelease	데이터 공개 수준
SDMX.Integrity.revision	데이터 개정 정책
SDMX.Integrity.notificationOfMethodologicalChange	조사방법 변경 공시
SDMX.Quality.dataSource	표제와 출판빈도
SDMX.Quality.crosschecksAndServiceability	유관 자료와의 호환성 등
SDMX.Methodology.definitions	방법론 개념 및 정의
SDMX.Methodology.scope	데이터 취급 범위
SDMX.Methodology.classifications	방법론적 분류 국제적 지침
SDMX.Methodology.recordingOfTransctions	방법론적 처리기록 회계와 예외 조항
SDMX.Methodology.natureOfTheBasicStatisticalData	기초 통계 데이터의 특성 수집절차
SDMX.Methodology.compilationPractice	통계 데이터 편집 및 가공
SDMX.Methodology.otherAspect	기타 요소 기타 주요 요소

하였다. 또한 매핑 결과에 따르면, 유일하게 통계청 내부 자료용 메타데이터 요소 가운데 '예산 · 인력 · 교육 · 홍보'와 '부록' 항목을 더블린 코어나 SDMX와 SDDS의 중복요소로 표현할 수 없었다. 이는 국제기준이 일반 이용자를 고려한 것이고, '예산 · 인력 · 교육 · 홍보' 요소는 통계청의 내부 이용자를 염두에 둔 것으로서 우리나라의 독특한 환경에 기인한 것으로 판단된다. 또한 앞에서 지적한 바와 같이 '부록'은 우리나라 조사보고서의 많은 부분이 인쇄형으로 발간되기 때문에 불가피하게 필요한 요소였다. 따라서 웹 기반의 국가통계 데이터를 공개할 경우에 이 항목에 대한 표준요소는 고려하지 않아도 무방하다고 판단하여 이를 제외하였다.[4] SDMX와 SDDS의 중복요소를 채택한 요소는 국제적인 통계자료의 교환 및 공유적 차원을 고려한 것이다. 더블린 코어의 경우는 국제수준의 자료공유와 관계없이 전통적으로 책자형 자료의 코딩에서 SDMX와 SDDS의 중복요소에 비해 상대적으로 적합한 형태라는 것과 국내 조사보고서의 초기 발간형태가 책자형에서 시작되었기 때문에 '목차'와 '머리말', '이용자를 위하여' 등과 같은 요소정보는 더블린 코어로 코딩하는 것이 바람직하다고 판단하여 표준 메타데이터 요소로 선정하였다. 최종적으로 국내 통계환경을 고려한 결과, SDMX와 SDDS의 중복요소와 더블린 코어로서 국내 통계조사보고서와 통계청 내부 메타데이터 요소의 대부분을 수용할 수 있었다.

5. 결 론

국가 통계 데이터는 하나의 국가에 대한 현황과 기록을 수치로 표현한 것으로서 해당 국가에 대한 종합적인 정보라 할 수 있다. 따라서 국가간 무역과 정치에 있어 통계 데이터의 중요성은 급증하고 있다. 왜냐하면 전 세계 모든 국가는 국가 통계를 정부차원에서 관리 및 통제하고, 해당 데이터의 국제적 공유 및 교환을 위한 의무를 갖고 있기 때문이다. 우리나라도 예외는 아니어서 국가 기본 통계 데이터를 OECD와 UN 등과 같은 국제기구에 통보하고, 국민의 알권리 증진을 위해 거의 모든 국가의 통계자료를 통계청 홈페이지를 통해 웹 상에 공개하고 있다.

한편 국가통계 데이터에 대한 요약의 수준과 표현방법은 각 주제별 혹은 국가상황의 특성에 따라 독특한 특성 때문에 통계 데이터 해석과 통합에 많은 제한점이 노정되고 있다. 이러한 제한점을 극복하기 위해서는 통계 메타데이터가 필요하다. 이러한 필요성에 따라 국제기구들이 주축이 되어 통계 데이터 교환용 부호화 포맷으로서 SDMX를 제안하고 있다. SDMX도 다른 국제 기준과 같이 여러 국가의 독특한 특성을 모두 수용하기 위해 기준의 범위와 내용이 일반적이며 범용적인 전개형태를 갖고 있다.

4) 현재 통계청에서 온라인 형태로 발간되는 자료의 경우에 부록정보는 제공하지 않고 있다.

우리나라의 국가통계 데이터도 SDMX와 SDDS의 중복요소로 완전하게 표현할 수 없기 때문에 우리나라 통계 데이터의 특성을 수용할 수 있는 확장 세트가 필요하다. 따라서 본 연구에서는 국가통계 데이터의 자유로운 활용과 공개를 위한 메타데이터 표준안으로서 SDMX와 SDDS의 중복요소와 함께 대표적인 메타데이터인 더블린 코어의 일부 요소를 확장하여 국가통계 메타데이터를 제안하였다.

확장의 기준은 국내 현황과 국제표준을 근거로 이루어졌다. 국내 현황 데이터의 수집은 1) 통계청 조사보고서의 형태정보와 2)통계청 내부용 확장 메타데이터 기준의 중첩도를 기준으로 통계 데이터로서 정보적 가치를 갖는 요소만을 메타데이터 요소로 선정하였다. 분석된 결과를 기준으로 통계 데이터의 국제표준인 SDMX와 SDDS를 적용하였다. 적용결과, SDMX와 SDDS의 중복요소를 기준으로 표현하기 어려운 항목은 더블린 코어의 한정어를 이용하여 국가 통계표준 메타데이터를 설계하였다. 이 표준은 '예산 · 인력 · 교육 · 홍보' 항목 이외에 국내 주요 통계조사보고서의 항목과 통계청 국가메타데이터 항목을 모두 표현할 수 있어 국내 통계 데이터의 국제적 공유와 호환을 위한 기준이 될 수 있도록 하였다. 그 결과, SDMX와 SDDS에서 중복된 29개의 항목을 채택하고, 더블린 코어에서 14개의 항목을 채택하여 최종적으로 43개로 이루어진 국가 통계표준 메타데이터를 완성하였다.

본 연구가 웹 기반의 국가 통계 데이터베이스 구축에 실제적으로 적용되기 위해서는 후속 연구로서 국가통계 메타데이터 표준 DTD 설계와 같은 연구가 수행되어야 할 것이다.

참고문헌

권용수. 1998. 『과학기술 통계 · 지표의 데이터베이스와 인터넷 유통 서비스 시스템 구축』. [서울] : 과학기술정책연구원. STEPI 정책연구보고서 98 22.

김태수. 더블린 코어(Dublin Core). [인용 2004. 8. 25]. <http://dewey.yonsei.ac.kr/metadata/DC.htm>.

박원환, 황조연. 2004. 통계자료의 비밀보호를 위한 익명화 방법들. 『통계연구』, 9(2) : 146 172.

이재창, 전명식, 김은석. 1997. 통계적 메타데이터의 역할과 표준화를 위한 추세. 『1997 춘계학술대회논문집』. [서울: 한국통계학회].

한국. 통계청. 2004. 『통계 메타 DB의 구축』. [대전] : 통계청. 한국정보관리학회 용역보고서.

Colledge, Michael, Fred Wensing, and Eden Brinkley. 1996. "Integrating Metatdata with Survey Development in a cai Environment". *1996 Annual Research Conference Proceedings*. US Census Bureau, [cited 2004. 8. 10].

〈http://www.census.gov/prod/2/gen/96arc/ xiiiwen s.pdf〉.

ECE, Eurostat, and OECD. 2004. "Practical Experience Towards Implementing SDMX at OECD Invited Paper". in *Meeting on the Management of Statistical Information Systems(MSIS)*, Geneva, May : 17 19.

Hustoft, Anne Gro, Jenny Linnerud, and Hans Viggo Sebo. 2004. "Quality and Metadata in Statistics Norway". *European Conference on Quality and Methodology in Official Statistics*(Q2004), May, Mainz : 24 26.

Lestina, Gregory J. Jr., William P. LaPlant Jr., Daniel W. Gillman, and Martin V. Appel. 1996. "Technical Development of the Proposed Statistical Metadata Standard". *U.S. Census Bureau*. (cited 2004. 7. 17). 〈http://www.census.gov/srd/www/metadata/ASA96TOC.HTML〉.

Pellegrino, Marco, and Denis Ward. 2004. " SDMX Metadata Common Vocabulary". *Statistical Data and Metadata Exchange*. April.

UN. 1999. " Information Systems Architecture For National and International Statistical Offices Guidelines and Recommendations". *Conference of European Statisticians Statistical Standards and Studies 51*. Geneva: UN.

UN. 2000. *Guidelines for Statistical Metadata on the Internet*. Geneva: UN.