XML에 기반한 국가 통계 메타데이터 표준화 연구

A study on statistical metadata of a Nation in XML base for standardization

하유정, 중앙대학교 문헌정보학과, yjh1021@hanmail.net
이두영, 중앙대학교 문헌정보학과, leety0521@hanmail.net

Yoo-Jung Ha, Graduate School of Library & Information Science, Chung-Ang University
Too-Young Lee, Dept. of Library and Information Science, Chung-Ang University

통계 데이터의 메타데이터에 대한 연구는 90년대 중반부터 각 나라에서 다양하게 이루어져 왔지만 아직 그 기준의 모호하다. 하지만 방대한 양의 데이터 관리와 자료의 활성화를 위한 방안으로 표준화를 위한 노력은 계속되고 있다. 여기서는 통계데이터를 위한 각 국가의 연구들을 조사하고 우리나라 통계청의 자료를 분석하여 통계 메타데이터 요소를 추출하였다. 본 연구는 이 요소들을 중심으로 XML 기반에서 DTD로 하나의 문서형태를 정의하여 우리나라 통계 데이터의 표준화를 위한 방안을 제시하고자 한다.

1. 서론

1.1 연구의 필요성 및 목적

급변하는 21세기에서 체계적이고 효과적으로 계획을 수립하고 정확한 정보를 얻기 위해서는 과거의 정보와 현재의 정보를 비교하는 것이 반드시 필요하게 되었다. 이런 시대의 요구 속에 정보를 수치로 정확하게 비교할 수 있는 통계자료의 중요성은 점차 증대하고 있다. 한편, 통계데이터에 대한 정량적이고 방대한 데이터의수집이 이루어지고 있기 때문에 통계관리의 필요성이 급성장하고 있다.

통계 데이터 국가 인프라의 완성에 따라 어디서나 정보접근이 가능하게 되었으나 통계처리의 어려움 때문에 활성화 되지 못하고 있다.

이런 현상의 주된 요인으로는 이용자들의 전문적인 통계 용어에 대한 이해 부족과 자료 구성의 통일성 부족으로 들 수 있다.

통계 메타데이터는 이런 통계 데이터의 문제점을 해소하고자 하였다.

일반적으로 메타데이터는 데이터를 기술해주는, 데이터 자체를 정리해주는 2차적인 개념이다. 통계 메타데이터는 통계 자료에 대한 데이터 수집과 관리를 지원한다. 또한 메타데이터의 요소로 통일된 데이터는 비전문가들의 자료 이해에 도움이 된다.

각 국은 이런 통계 메타데이터를 구축하기 위해 많은 노력을 하고 있다. 그러나 각 나라의 통계 메타데이터에 대한 정의만으로는 국제 정보를 공유하기 어렵기 때문에 국제적으로 협력하고 있다. 이런 노력의 일환으로 더블린코어 (Dublin Core)와 XML (eXtensible Markup Language)을 들 수 있다. 그 밖에 유럽에서는 양질의 정보를 제공하고 데이터를 공유하기 위한 목적으로, 1994년부터 EUROSTAT (Statistical Office of the

European Communities)를 운영중이다. 이곳에서는 정기적으로 시대의 변화에 맞게 통계 메타데이터를 정의하고 발전시키고 있다.

우리나라 통계청에서는 KOSIS (Korean Statistical Information System)라는 통계정보 시스템을 통해 데이터를 효과적으로 제공하고 관리하기 위해 노력한다. 최근에는 현재의 이질적인 통계 데이터의 표준화를 위한 통계 메타데이터 개발과 데이터의 활성화에 주안점을 두고 있다.

이런 표준화를 위한 연구의 핵심은 공통 문서 형태의 정의이다.

표준화된 문서 형태는 다음과 같이 효율적인 데이터 관리에 도움이 된다. 첫째, 이질적이고 분산된 자료들의 데이터베이스 구축을 위한 시간과 노력을 절약해 준다. 둘째, 전반적인 데이터의 관리가 용이하고, 데이터를 일괄적으로 처리해준다. 셋째, 이용자들은 하나의 표준 문서형태로 인해 다른 모든 데이터를 이해할 수 있다.

이런 표준화 문서 형태를 위한 가장 최선의 방법은 웹상에서 문서 관리에 적합한 XML의 이용이다. 여기서 DTD(Document Type Definition)는 문서의 형태 정의를 위한 도구가 이다. 즉, 통계 메타데이터는 DTD 형식으로 표준화된 하나의 문서 형태로 정의 될 수 있다.

통계 메타데이터는 이 DTD의 구성요소이기 때문에 DTD 설계를 위해 필수적이다. 그리고 기존의 DTD의 요소로는 통계 데이터만의 특징을 표현하기 어렵기 때문에 통계 자료만을 위한 DTD의 설계가 필요하다. 본 연구에서는 우리나라 통계 데이터의 문서 표준화를 위한 통계메타데이터의 추출과 DTD 구축에 그 목적을 둔다.

1.2 연구방법

본 연구는 우선 통계 메타데이터에 대한 개념을 정리한다. 그 다음 각 국의 통계 데이터 활성화 방안을 조사하고 현황을 파악한다. 이에 대상이 되는 국가는 미국, 캐나다, 호주가 주를 이룬다.

우리나라의 경우 통계청에서 생산하는 인쇄형태 자료를 중심으로 분석한다. 이것은 자료의 물리적 형태와 그 자료의 핵심 내용인 조사보고서를 비교 분석함으로 일어난다. 여기에 나타난 각 자료의 공통의 메타요소는 외형적인 부분과 내형적인 부분으로 나누어서 추출된다.

정리된 통계데이터의 메타요소를 일반적인 데이터의 DTD 요소와 비교하여 필요한 요소를 선택한다. 그 다음 통계 데이터만을 위한 DTD 요소를 정의한다. 마지막으로 이것을 XML 기반에서 우리나라 통계 자료의 표준화를 위한 통계자료의 외형적 DTD 정의와 내용적 DTD의 정의로 나누어 구축한다.

2. 용어의 개념정의

1) 통계데이터(statistical data)

UN산하의 통계위원회에서 발행한 통계메타데이터 용어집에서는 통계데이터를 통계조사 또는 통계적 데이터를 처리하는 과정에서 통계에 의해 수집·생산되는 데이터로 정의하고 있다.

2) 통계 메타데이터(statistical metadata)

UN산하의 통계위원회에서 발행한 통계메타데이터 용어집에서는 통계 메타데이터를 '통계데이터에 대해 기술하는 데이터'로 정의하고 있다. SDMX(Statistical Date and Metadata Exchange)에서 발행한 메타데이터 공통 용어집에서는 이를 더욱 요약하여 '통계데이터에 대한 데이터'로 정의하고 있다. 여기서는 '데이터에 대한 데이터'라는 정의가 갖는 모호함을 없애기 위해 UN 통계위원회의

견해를 차용하였다. 본 연구에서는 통계 메타데이터를 통계데이터에 대해 정의하거나 기술하는 데이터를 문서화하는 과정으로 정의한다. 실질적으로 통계 메타 데이터는 통계데이터에 숨어 있는 (behind)데이터로 설명하고 있다. (한국정보처리학회, 2004)

3. 통계적 메타데이터의 특징

통계적 데이터는 다음과 같이 두 가지로 구분 할 수 있다.

Microdata - 개인, 가구, 단체 같은 모집단의 개체에 대한 특성을 측정한 자료로 실험이나 조사 센서스에 의해 수집된 데이터.
Macrodata - 빈도, 평균, 계산과 같은 것으로 마이크로 데이터로부터 얻어낸 마이크로 데이터를 가공하여 데이터로써 상대적으로 많은 통계정보를 갖고 있다.

또한 통계적 메타데이터를 구성요소 별 혹은 단계로 나누면 다음과 같다.(Census Bureau, 1999)

Systems - 위치, 레코드 레이아웃, 데이터베이스 스키마, 매체의 종류, 사이즈 등과 같은, 응용 데이터 세트의 물리적인 특성에 대한 정보
Application - 표본추출설계, 질문, 소프트웨어, 변수 정의, 전문적인 편집 등과 같이 응용 생산물과 과정에 대한 정보
Administrative- 예산, 비용, 계획 등과 같은 관리 요소

이런 통계적 메타데이터는 통계정보의 잠재적인 이용자들을 돕기 위해 그리고 자동통합 관리 시스템을 통한 통계 조사의 계획과 설계와 운영 등의 전체과정을 지원하기 위해 구축된다.

4. 통계 데이터 활용을 위한 연구 현황

1) 해외

캐나다의 경우, 통계 프로그램의 메타정보를 위한 중앙 저장고인 Integrated MetaData Base (IMDB)가 있다. 이것의 하나인 Phase 3 는 국제 표준인 ISO 11179용어를 따른다. 이것은 데이터를 더 전개시키기 위한 규정을 만들기 위해 일치, 분류, 정의하는 것에 대한 표준을 선언한다. 또 표준화를 위해, 용어를 통일하여 Phase 3 모델의 데이터 요소, 성분, 관계에 대한 가이드라인을 설정해준다, 하지만 용어의 표준화만으로는 통계의 다양성이나 변이를 수용하는 한계를 알고 자체적으로 융통성 있고 변화가 자유로운 그들만의 메타데이터를 발전시키고 있다.

미국의 경우, 미국 통계국에서는 통계데이터를 이용자중심의 검색 및 관리 프로그램으로 만들어 제공하고 있다. Automated Reference Rack(ARRK)은 로터스 Smart Text를 이용해 설계된 하이퍼텍스트 기반 시스템이다. 통계국직원들은 ARRK를 이용하여 이용 가능한 파일에 대해 짧게 기술하고, 일반사용자들은 이를 통해 출간된 자료에 대한 검색을 쉽게 할 수 있도록 구축되어 있다. 이때 파일에 기술된 내용은 주문정보, 비용, 크기, 저장매체, 주체, 적용범위 등이다.

2) 국내

통계청에서 운영중인 통계정보시스템(KOSIS)은 국내 주요통계와 UN 및 IMF의 국제비교 통계정보 등 약 994만 계열의 주요 통계정보를 수록한다. 자료의 유형별 검색과 주제별, 지역별, 통계작성기관별, 색인어별 검색이 가능하고 시계열 검색을 할 수 있게 한

다. 보건복지부에서는 통계정보시스템을 구축하여 보건복지부와 관련기관에서 생성한 통계들을 실시간으로 제공한다. 한국은행은 주제분류와 통계검색 그리고 시계열 검색과 횡단검색을 제공하고 있다. 또한 단일자료구분에 각 항목들을 가로 혹은 세로로 배열하여 검색하는 방법과 주요경제지표에 따른 통계표를 제공한다.

5. 우리나라 통계청 메타데이터 분석 및 추출

통계청의 메타데이터는 크게 발간자료의 물리적 구조정보와 조사표를 통한 내용적 구조정보로 나누어 추출하였다.

1) 외형적 메타구조

통계청에서 생산되고 있는 자료의 형태는 조사의 규모나 특성을 고려하여 조사표와 산업통계데이터, 수주통계데이터, 인구조사 데이터로 구분할 수 있다. 이상의 자료를 다음과 같이 외형적 특성을 파악하기 위해 실제 자료를 조사하여 분석하였다.

다음 예는 2002년 12월중 도·소매업 판매액지수를 비롯하여 최근 3년간의 해당 조사보고서에 등재된 외형적 요소를 분석한 것이다.

표제지를 비롯하여 외형적 특성은 크게 표제지와 머리말, 이용자를 위하여(일러두기), 차례(목차), 조사개요, 조사결과요약(개요), 통계표, 부록으로 구분된다. 또한 이 항들은 필요에 따라 하위항목은 다음과 같다.

- 표제지 : 등록번호, 년도, 제목, 항목, 조사범위, 발행기관, 발행일 등
- 머리말
- 이용자를 위하여 : 부호설명, 연락처
- 제목 (목차)
- 조사개요 : 조사목적, 연혁, 조사의 법적 근거, 조사 기준시점 및 조사기간, 조사범위 및 대상, 조사체계, 조사항목, 조사방법, 조사수행조직, 집계 및 공표, 용어해설, 표본설계
- 조사결과요약(개요)
- 통계표
- 부록 : 조사표 실물, 통계청발간간행물

2)내용적 메타구조

각 조사보고서의 내용적 속성을 측정하기 위해 각 조사보고서에 있는 통계조사표를 분석하였다. 분석은 세 개의 부분으로 나누어진다. 산업관련, 수주통계보고서, 인구관련 통계조사표 이다.

이들 자료의 조사표를 비교해보면, 산업관련과 수주통계자료에서는 거의 공통적으로 '사업체 명 및 소재지', '경영조직 및 조직형태'. '자본금'.'종업원 수 및 연간급여액', '재고액'. '연간 출하액 및 수입액', '납부부가가치세', '재산' 과 같은 항목이 필수적으로 들어가 있다.

인구(가계관련)조사에서는 '성명', '가구주와의 관계'. '성별'. '생년월일', '나이', '교육정도', '출생지'. '혼인사항' 같은 사항이 공통적으로 들어가 있다.

각 자료의 기타요소는 그 조사의 목적에 맞게 항목들을 산출하여 조사하고 있다.

이런 통계청 자료의 항목들은 빈도수가 높은 항목과 각각의 특성이 담긴 고유의 항목과 변수항목들도 구성된다.

본 연구에서는 빈도수가 높은 주요 항목은 필수적인 메타데이터 요소가 된다. 변수항목들 역시 자료의 중요한 식별기준이 되기 때문에 변수항목으로서 메타데이터의 요소가 된다.

6. XML 에 기반한 DTD 설계

XML은 1996년 W3C(World Wide Web Consortium)의 XML Working Group에서 제안한 웹상에서 구조화된 문서를 전송 가능하도록 설계된 표준화된 텍스트 형식이다. XML은 기존의 HTML(Hyper Text)과는 달리 웹상에서의 정보 제공자에게 그들이 제공하는 정보에 대한 자신의 구조를 정의 할 수 있도록 하고, SGML(Standard Generalized Markup Language)의 불필요한 부분을 축소하여 구현하기 쉽게 만들어졌다. 웹상에서 SGML의 기능을 구현하는 것을 그 목적이 있다.

DTD는 문서구조를 정의하는 것으로 문서와는 독립적으로 저장 될 수 있다.

DTD는 또한 네 개의 섹션을 가지고 있다. 기술용 메타데이터 섹션은 외부의 메타데이터 레코드를 지시하거나 또는 임베드된 기술용 메타데이터를 수록할 수 있다. 관리용 메타데이터 섹션은 세 가지 유형의 데이터, 즉 파일 작성 및 물리적 특성에 관한 테크니컬 메타데이터, 지적 재산권 정보, 디지털 객체의 원본 소스에 관한 정보를 정의 한다. 구조용 메타데이터 섹션은 파일들을 구조화된 문서의 일부로서 논리적 위치로 조직화 한다. 마지막으로 "파일 인벤토리"(file inventory)라는 섹션은 아카이브 객체의 특정 버전(예를 들면 JPEG 또는 썸네일 버전)을 위한 모든 파일들을 그룹화 한다.(Prisclla Caplan, 2003)

DTD에는 엘리먼트(Element), 선언, 엔티티(Entity) 선언, 속성(attribute)선언이 포함된다. DTD 명세 안의 한 엘리먼트는 그것의 이름에 의해서 기술되고, 하나의 엘리트먼트가 구성될 수 있는 방법에 의해서 기술된다. 예를 들면 저널 논문 형태의 DTD 엘리먼트는 제목 정보, 요약, 내용, 참고문헌으로 구성될 수 있다. DTD 안에서 저널 논문의 엘리먼트 정의는 다음과 같다.

<! ELEMENT Journalpaper -- (TitleInfo, Abstract, Contents, Reference) >
<! ELEMENT TitleInfo -- (Authors, Affiliation, Address)>
<! ELEMENT Contents -- (Sections)>
<! ELEMEMT Sections -- (Paragraphs, Figures, Tables)>

문서 안에 모든 엘리먼트들은 DTD 안에서 완전하게 정의될 수 있다. 엘리트먼트의 추가적인 특성은 속성들의 의미에 의해 기술된다. 속성들은 요소들의 특징을 나타낸다.

위에서 본 것과 같이 일반적인 자료는 더블린코어에서 정의해준 메타데이터의 서지적 사항의 정의와 유사하기 때문에 그것을 DTD에 그대로 이용할 수 있다. 하지만 통계자료의 경우 통계자료의 특성이 담겨 있는 다른 형태의 DTD가 필요하다.

예를 들어 외형적인 메타요소의 정의에 있어, 일반적인 인쇄 자료는 표제지에 년도가 기재된 경우가 드물다. 하지만 통계자료의 경우 자료의 특성상 제목에 반드시 년도가 필수적 이다. 제목에 있는 년도와 표제지에 있는 발행일 또한 다르기 때문에 각각에 대한 정의가 필요하다. 또한 표제지 부분의 통계 자료의 고유의 등록번호 역시 일반적인 DTD의 형태에 맞지 않다.

무엇보다 통계자료의 내용적인 메타데이터 요소 정의를 위한 DTD에는 통계자료에 필수적인 조사표 부분이 요구된다. 이런 조사표에 대한 정의는 산업부분과, 수지부분, 인구(가계)부분 등 그 자료의 주제 구분에 따라 다른 정의가 설계되어야 한다.

일반적인 문서형태 정의를 위한 요소에는 이런 차이점에 대한 정의가 부족하기 때문에 통계 데이터만을 위한 DTD의 구축이 요구된다.

7. 결 론

본 연구에서는 우리의 통계자료에 맞는 통계 메타데이터를 추출하였다.

통계자료의 특성상 두 가지 형태의 데이터 형태 정의가 필요하다. 하나는 물리적 형태에 대한 정의이고 하나는 조사표 형태에 대한 정의이다.

두 가지 형태의 정의 모두 공통적으로 들어가는 필수요소와 변수 항목을 위한 가변적인 형태의 정의가 필요하다. 또한 조사의 주제에 따라 메타데이터의 구조와 항목이 다르기 때문에 주제별로 나뉜 각각의 정의가 요구된다.

하지만 일반적인 문서를 위한 DTD는 이런 조건에 부합하지 않기 때문에 통계 데이터만을 위한 DTD 설계가 필요하다.

본 연구는 기존의 DTD자료와 더블린코어의 정의를 참고로 통계 자료의 외형적 메타데이터 요소를 정의한다. 자료의 내용적인 데이터의 요소는 독자적으로 정의한다. 또한 이것은 통계 자료만의 특성을 표현할 수 있는 DTD로 구현될 것이다.

이런 DTD의 구축은 이질적이고 분산된 통계정보를 통합된 하나의 형태로 통일 시켜 주고 통계데이터 관리를 위한 비용과 시간을 절약해 줄 수 있을 것이다. 더 나아가 통계 데이터의 표준화는 이용자들의 통계데이터 이용에도 도움이 될 것이라 기대된다.

참고문헌

이재창, 전명식, 김은석, 통계적 메타데이터의 역할과 표준화를 위한 추세, 고려대학교, 한국통계학회, 1997

한국전산원, 공공부문의 DTD 개발지침, 한국전산원, 1998

김태순, XML 활용에 의한 EDI 구현에 관한 연구, 한양대학교, 2000

신동일, 신동규 공저, 멀티미디어 데이터베이스, 인터비젼, 2002

Prisclla Caplan 저, 오동근역, 메타데이터의 이해, 태일사, 2004

한국정보관리학회, 통계메타DB의 구축, 통계청, 2004

Daniel W. Gillman, Martin V. Appel, Statistical Metadata Research at the Census Bureau, Census Bureau

Paul Johanis, Brad Brooks, Tim Dunstan, Statistics Canada's Implementation of The Data Element Model, Statistics Canada, 2003