KORMARC의 DTD 및 변환프로그램 개발 연구

A Study on the Development of a KORMARC DTD and Conversion Programs

김성은, 김태수 (연세대학교 문헌정보학과)
Seong-Eun Kim, Tae-Soo Kim (Dept. of Library and Information Science, Yonsei Univ.)

목록이 새로운 정보환경에서도 효과적인 정보서비스에 기여하기 위해서는 그 내용과 형식 모두에 대한 개선이 절실하게 요구된다. 특히 지속적인 내용의 개선사항을 반영하고 네트워크 환경에서의 융통성있는 사용을 위해서는 새로운 목록데이터 형식이 필요하다. 이에 본 연구에서는 MARC 포맷의 개선을 위해 메타언어인 SGML의 응용 타당성을 제시하고, 이를 바탕으로 KORMARC에 대한 DTD 및 KORMARC/SGML 두 형식간 양방향 변환프로그램을 개발하였다.

1. 서론

정보자원의 양적 증가와 그 유형의 다양화, 그리고 네트워크의 확대 및 보편화는 정보조직(information organization)의 중요성을 더욱 부각시켰고, 이에 따라 정보의 조직에 필수적인 메타데이터에 대한 연구가 각 분야에서 활발하게 이루어지고 있다. 도서관 분야의 메타데이터인 목록도 이러한 정보환경에 부합하는 효과적인 정보서비스에 기여할 수 있어야 하며 이를 위해서는 목록에 대한 새로운 시각의 접근방법이 요구된다.

현재 사용되고 있는 MARC는 목록규칙과 관련된 그 내용뿐만 아니라 데이터기술언어라는 측면에서 볼 때 마크업이나 레코드구조 등 그 형식에 있어서도 많은 문제를 갖는 것으로 나타나고 있다. 다양한 정보자원을 기술하고 네트워크 환경의 여러 응용분야에서 사용되기에는 한계를 드러내게 된 것이다. 이에 본 연구에서는 MARC 포맷의 개선을 위한 방법론으로서 MARC와 같은 데이터기술언어를 정의할 수 있는 메타언어 SGML의 응용을 제안하고자 한다.

2. 이론적 배경

2.1 메타데이터

메타데이터란 데이터에 관한 데이터, 즉 데이터 혹은 정보의 여러 속성을 기술해 주는 데이터이다. 정보의 체계적인 조직에 있어 그 의미가 크며 정보자원의 식별 및 소재 파악, 내용 기술 등의 역할을 하게 된다. 정보자원의 유형과 접근방법이 다양해짐에 따라 메타데이터도 주제와 응용분야, 기술대상 정보자원과의 조직방법, 기술의 상세성 및 구조화 수준, 그리고 작성 주체 등에 있어서 매우 다양한 양상을 띠게 되었다. 현재의 정보환경은 적절한 메타데이터를 기반으로 하여 이용자가 적합한 정보자원을 탐색하고 발견할 수 있도록 하는 체계 및 기법의 개발이 더욱 중요하게 되었다.

다양한 메타데이터의 개발과 함께 이들간의 호환성 확보와 통합환경 구축에 대한 연구도 활발하다. 이러한 시도는 MARC와 같은 단

일 포맷을 이용한 통합 시도와 여러 메타데이터를 하나의 구조적인 체계 안에서 운용하고자 하는 시도로 나누어 볼 수 있다.

특히 MARC를 이용하는 경우는 생성된 레코드를 기존 온라인열람목록 등을 통해 검색할 수 있기 때문에 이미 구축되어 있는 MARC 기반의 정보하부구조를 그대로 사용할 수 있음을 의미한다. 반대로 여러 메타데이터 체계의 다양성과 그 필요성을 인정하고 이들을 하나의 통합적인 틀 안에서 운용하고자 하는 추세도 강하다. 이를 위해서는 각 메타데이터간 상호작용이나 데이터의 코딩, 메타데이터 유형의 등록, 네트워크의 효율성, 데이터의 교환과 검색을 위한 접근 등의 문제가 선결되어야 한다.

2.2 SGML을 이용한 메타데이터 기술

메타데이터에 대한 연구 경향 중 두드러지는 것은 SGML의 응용이다. SGML은 첫째, SGML 문헌의 구성방식에 의해 플랫폼 및 응용분야와의 독립성이 확보되고 따라서 이식성, 내구성, 호환성, 확장성, 융통성 등을 얻을 수 있다. 둘째, 구조정보의 표현력이 뛰어나 MARC와 같이 그 기술의 상세성 및 구조화 수준이 높은 메타데이터에 있어 효과적이다. 셋째, SGML로 된 정보자원이 증가함에 따라 메타데이터도 SGML로 기술하게 되면 메타데이터와 원정보자원의 통합이 가능하게 됨으로써 이용자가 해당 정보자원으로 직접 연결하는 일이 훨씬 수월해진다.

이와 같은 여러 가지 장점으로 인해 통합환경의 구축 및 네트워크 환경에서의 원활한 사용을 위한 최적의 메타데이터 기술언어로서 SGML이 제시되고 있는 것이다.

2.3 SGML과 MARC

SGML은 도서관의 모든 기능 및 서비스에 적용되고 있으며 크게 전문데이터 영역과 서지데이터 영역의 두가지 흐름으로 구분할 수 있다. 서지데이터 영역인 목록에 대해서는 특히 MARC에 대해 다양한 시도가 이루어지고 있는데 이는 서지정보의 기술구조인 MARC가 기술적 마크업언어이면서 목록규칙과 같은 엄밀한 논리적 구조를 갖고 있고 또 메타데이터 중 비교적 그 기술 내용이 상세하고 구조가 정교해 SGML의 적용이 용이하면서도 효과적이기 때문이다.

그러나 무엇보다도 MARC가 드러내고 있는 포맷의 경직성이 SGML이라는 대안의 사용을 모색하게 한 주된 원인이라 할 수 있을 것이다. ISO 2709에 의해 규정된 현 MARC 포맷은 1970년대 초반의 컴퓨터 환경을 반영한 것으로서 데이터 교환 및 저장매체 기술의 발전을 반영할 수 있도록 개선되지 못하고 있어 새로운 데이터요소의 추가나 기존 포맷의 수정 등을 매우 어렵게 만들고 있다. 특히 네트워크 환경에서 그 중요성이 커지고 있는 분립적 비서지정보의 구조화나 계층적 분석수준의 다양화, 그리고 한 저작과 관련된 여러 판의 식별 및 연결 등을 해결하는 데 있어 많은 문제점을 노출하고 있다.

따라서 새로운 유형의 정보자원을 처리하고 웹과 같은 정보환경에서도 효과적으로 서비스될 수 있으며, 다양한 표준을 따르는 여러 메타데이터 교환의 매개체로서, 그리고 포괄적인 통합환경 안에서의 구성요소로서 그 역할을 하기 위해서는 데이터 내용 자체의 개선사항을 지속적으로 반영하면서도 동시에 도서관외의 정보환경에도 통합될 수 있도록 융통성있는 포맷이 필요하게 된 것이다.

3. KORMARC DTD

3.1 개발원칙

SGML의 응용에 있어 가장 기본적인 것은 표준적인 DTD의 개발이라 할 수 있다. 따라서 본 연구에서는 USMARC DTD 초안을 참조하여 단행본용 KORMARC에 대한 DTD를 설계하였다. DTD를 이용한 MARC 형식과 SGML

형식간의 양방향 변환과정에서 정보의 손실이 없어야 한다는 점과, DTD의 내용이 현 MARC 표준을 따르고 그 유지, 보수가 병행되어야 한다는 점을 기본 원칙으로 하였다. 단행본용 KORMARC(KS C 5867)가 표준으로 채택되어 사용되고 있고 기존 데이터의 소급변환을 고려해야 하므로 쓰이지 않거나 유용하지 못한 데이터필드라 하더라도 그대로 DTD에서 정의하였으며 새로운 내용의 추가나 수정, 보완은 고려하지 않았다.

3.2 구성

전체 KORMARC DTD를 모두 16개의 DTD 파일로 나누어 설계하고 매개변수엔티티 참조기법을 통해 연결하였다. DTD의 골격을 형성하는 기본 DTD 파일('*kormarc.dtd*')과 한국문헌자동화목록형식에서 제공하는 필드그룹화 기준에 따른 14개 DTD 파일, 그리고 엔티티 파일 등으로 구성된다. '*kormarc.dtd*' 파일에서는 나머지 15개 DTD 파일에 대한 외부엔티티를 선언하고 KORMARC DTD를 구성하는 최상위요소들에 대해 정의한 후 각각에 대해 선언된 엔티티를 참조함으로써 하위계층요소에 대한 DTD 파일로 연결될 수 있도록 하였다.

```
<!DOCTYPE Kormarc [

<!ENTITY  % Entity-Leader              SYSTEM  "c:\kse\kormarc-dtd\leader.dtd">
<!ENTITY  % Entity-Control-Fields      SYSTEM  "c:\kse\kormarc-dtd\ctrl_fld.dtd">
<!ENTITY  % Entity-Numbers-And-Codes   SYSTEM  "c:\kse\kormarc-dtd\num_code.dtd">
                                               :
<!ENTITY  % Entity-Series-Added-Entry  SYSTEM  "c:\kse\kormarc-dtd\s_added.dtd">
<!ENTITY  % Entity-Others              SYSTEM  "c:\kse\kormarc-dtd\others.dtd">
<!ENTITY  % Entity-Entities            SYSTEM  "c:\kse\kormarc-dtd\entities.dtd">

%Entity-Entities;

<!ELEMENT Kormarc - - (Leader, Control-Fields, Numbers-And-Codes, Main-Entry?,
                       Title-And-Title-Related, Edition-Imprint-Etc?,
                       Physical-Description, Series-Statement?, Notes?,
                       Subject-Access?, Added-Entry?, Linking-Entry?,
                       Series-Added-Entry?, Others?)>
<!ATTLIST Kormarc  type  (book|serial|mss-book)  #REQUIRED
                   id    ID                      #IMPLIED>

<!-- ================================== 리더 ================================== -->
<!ELEMENT Leader - - (ldr05, ldr06, ldr07, ldr08-09, ldr17, ldr18, ldr19)>
<!ATTLIST Leader  name  CDATA  #FIXED  "리더">

%Entity-Leader;

<!-- ================================== 00X 제어필드 ================================== -->
<!ELEMENT Control-Fields - - (kmrc001, kmrc005, kmrc006*, kmrc007*, kmrc008)>
<!ATTLIST Control-Fields  name  CDATA  #FIXED  "제어필드">

%Entity-Control-Fields;

<!-- ================================== 01X~09X 부호필드 ================================== -->
▷
```

3.3 세부설계사항

필드그룹 14개, 즉 00X 제어필드, 01X~09X 부호필드, 1XX 기본표목, 20X~24X 서명 및 서명관련사항, 250~29X 판차, 발행 등 사항, 300 형태사항, 4XX 총서사항, 5XX 주기사항, 6XX 주제명부출표목, 70X~740 부출표목, 76X~79X 연관저록, 830 총서부출표목, 850~890 기타 등을 최상위계층 요소(element)로 하여 KORMARC 표시기호로 구분되는 기본필드에 대한 요소 89개, KORMARC 식별기호에 의해 구분되는 하위필드에 대한 요소 232개, 그리고 리더를 비롯한 고정길이필드의 각 자수위치에 대한 요소 54개 등 총 389개 요소로 구성하였다. 요소의 출현규칙은 한국문헌자동화목록형식에서 제시하고 있는 각 필드의 적용수준과 반복사용여부에 대한 내용을 토대로 한다.

DTD에 사용된 속성(attribute)은 모두 네 가지로 KORMARC가 대상으로 하는 자료유형에 대한 'type' 속성, 각 필드요소에 대한 한글이름을 나타내는 'name' 속성, 지시기호를 표현하는 'ind1/ind2' 속성, 그리고 고정길이필드의 경우 해당 필드가 가질 수 있는 값을 가리키는 'value' 속성 등이다.

엔티티(entity)의 사용은 모듈화된 DTD 설계원칙에 입각하여 파일참조를 위한 외부엔티티와, 개인명, 단체명, 회의명, 통일서명, 총서사항, 주제명부출표목 등을 구성하는 하위필드처럼 DTD 상에서 반복적으로 출현하는 내용을 효율적으로 처리하기 위한 매개변수엔티티 등 두가지 유형을 사용하였다. 본 연구에서는 특수문자나 외국문자 코딩을 위한 엔티티는 고려하지 않았다.

4. KORMARC/SGML 변환프로그램

정의된 KORMARC DTD에 의거하여 정보의 손실없이 양방향으로 목록데이터의 형식을 변환할 수 있고 다량의 데이터 변환작업을 효율적으로 수행하여 두 형식 데이터에 대해 모

두 전거제어와 같은 목록업무를 병행할 수 있도록 하는 데 목적을 두었다. 프로그래밍언어인 비주얼베이직(Visual Basic) 4.0을 이용하여 윈도95 환경의 'kormarc2sgml', 'sgml2kormarc' 두 프로그램을 통합 구성하였으며, 표준 DTD의 개발이 완료될 때까지 변경되는 내용을 계속해서 반영할 수 있도록 각 필드에 대한 처리나 여러 수행기능에 대해 가능한 한 독립적인 모듈을 작성하였다. 본 프로그램은 두 형식간의 변환에 중점을 두고 있으므로 데이터를 파일에 효율적으로 저장하는 방법이나 데이터베이스의 구축 등은 고려하지 않았으며 변환결과는 일반 텍스트파일로 저장된다.

kormarc2sgml의 입력데이터는 KORMARC 표준에 따른 목록레코드의 원시데이터를 사용하며, sgml2kormarc의 입력데이터는 정의된 DTD에 따라 레코드 시작태그인 '<Kormarc>'와 종료태그인 '</Kormarc>'로 구분된다. 본 연구에서는 kormarc2sgml 프로그램을 통해 변환이 수행된 결과데이터를 sgml2kormarc의 입력데이터로 사용하였다. 다음은 이러한 SGML 형식 목록데이터의 예이다.

```
<!DOCTYPE Kormarc SYSTEM "c:\kse\Kormarc-dtd\kormarc.dtd">
<Kormarc type="book">
<Leader name="리더">
<ldr05 name="레코드상태" value="n">
<ldr06 name="레코드형태" value="a">
<ldr07 name="서지수준" value="m">
<ldr08-09 name="빈칸">
<ldr17 name="입력수준" value="blank">
<ldr18 name="목록기술형식" value="k">
<ldr19 name="연관레코드조건" value="blank">
</Leader>
<Control-Fields name="제어필드">
<kmrc001 name="제어번호">KM0199318004
<kmrc005 name="최종처리일시">19941010105753
<kmrc008 name="부호화정보필드">
<fln0-5 name="입력일자">940501
<fln6 name="발행년유형" value="s">
<fln7-10 name="발행년1">1992
<fln11-14 name="발행년2">
<fln15-17 name="발행국명">ulk
```

변환방식은 일괄처리와 일대일처리를 모두 지원한다. 일괄처리에서는 변환대상이 되는 여러 개의 파일들을 선택하여 이들을 한번에 변환할 수 있다. 일대일처리에서는 입력데이터 및 결과데이터의 원시데이터와 함께 해당 형식의 구조정보를 추출하여 도식화한 내용을 함께 출력해 준다. 또한 kormarc2sgml의 경우 변환시 해당 레코드에 출현한 필드그룹을 인식하여 사용자가 변환대상 그룹을 선택할 수 있도록 하였는데 이는 목록데이터와 다른 메타데이터간에 교환이 늘어나고 또 데이터의 공유라든가 모듈화된 메타데이터 작성이 활발해질 것이므로 필요에 따라 필드를 선택하도록 하는 것이 바람직하리라 판단되었기 때문이다. 일대일처리에서는 간단한 편집기능을 이용하여 데이터의 수정도 가능하다.

본 프로그램은 손실없는 변환과 효율적 변환이라는 두가지 기본 기능과 함께 구조정보 추출에 의해 데이터베이스 관련 시스템과의 연동이 용이하므로 SGML 시스템이나 MARC 시스템으로 데이터를 손쉽게 이식할 수 있다는 데에서 그 의의를 찾을 수 있다.

5. 결론

목록데이터의 형식 변환은 기존에 구축해 놓은 대량의 데이터를 활용하는 것이므로 새로운 정보환경에서 사용될 내용물의 효율적인 생성이라는 측면에서 그 중요성이 크다. 또한 양방향 변환을 통해 SGML 기반 시스템과 MARC 기반 시스템간의 호환성을 확보함으로써 총체적인 변화가 가져올 수 있는 여러 문제들을 방지하고 기존 도서관의 기법 및 절차와 형식을 바탕으로 점진적이고 안정적인 방법으로 새로운 정보환경을 구축해 갈 수 있다. 그리고 SGML 형식의 목록데이터 작성을 통해 다양한 메타데이터의 통합이나 원정보자원과의 통합 등 포괄적 통합환경의 구축을 앞당길 수 있으며 그 과정에서 목록이 주도적인 역할을 할 수 있는 것이다. 국가디지털도서관의 구축이라는 중요한 과제를 안고 있는 이 시점에서 그 토대가 되는 KORMARC를 새로운 정보환경에 적응시킬 수 있도록 내용과 형식 모두에 대한 끊임없는 연구가 필요하다.