

한국문헌자동화목록형식(KORMARC)의 XML 변환에 관한 연구

김수형[†], 이경현[‡]

[†] 부경대학교 전산정보학과, [‡] 부경대학교 전자컴퓨터정보통신공학부

A Study on Conversion of KORMARC(Korean Machine Readable Cataloging) into XML

Soo-Hyoung Kim[†] and Kyung-Hyune Rhee[‡]

[†] Dept. of Computer & Information Science, PKNU

[‡] Division of Electronic, Computer and TeleCommunication Engineering, PKNU

요약

인터넷에서의 정보자원의 증가에 따라 정보자원관리 기법의 관심이 높아지고 있는 가운데 기존의 HTML과 SGML의 단점을 보완한 구조화된 포맷인 XML의 응용에 대한 연구가 활발해지고 있다. 본 논문에서는 전통적인 도서관에서 서지레코드로 널리 사용되고 있는 구조화된 포맷인 MARC를 대상으로 현황과 문제점을 살펴보고, MARC를 XML로 변환하는 데 따른 기대효과를 고찰한다. 또한, 우리나라 MARC 포맷인 KORMARC을 XML로 변환하기 위하여 KORMARC에 대한 DTD를 설계한다.

1. 서론

인터넷 이용의 확산은 네트워크상의 정보자원의 폭발적인 증가를 가져왔으나 이를 효율적으로 이용하는 방법론은 아직 초기단계에 머물고 있다. 인터넷, 특히 WWW의 정보자원의 대부분을 차지하고 있는 HTML(Hyper Text Markup Language)은 애초에 정보 교환용 목적으로 설계된 포맷이 아니기 때문에 구조화된 정보검색과 교환이 어렵고, 다양한 멀티미디어 타입을 제공하지 못하며, 정교한 페이지 형태를 제어하기 어렵다. 그러나 1998년 XML(eXtensible Markup Language) 1.0이 제정되면서 HTML의 단순성과 기존 SGM(L(Standard Generalized Markup Language)의 복잡함에서 야기되는 문제점들을 동시에 해결할 수 있게 되었다. 즉, XML은 현재 구조화된 포맷의 현실적인 표준이라 할 수 있다.

본 논문에서는 가장 오래된 구조화된 포맷으로서 1960년대부터 전통적인 도서관에서 사용하고 있는 MA

RC(Machine Readable Cataloging) 포맷과 우리나라의 KORMARC(Korean MARC)의 문제점을 살펴보고 그 해결방안으로서 MARC 포맷을 XML로 변환하기 위한 필요성과 기대효과를 고찰하며, KORMARC의 XML 변환을 위한 사전단계로서 KORMARC의 DTD(Document Type Definition)를 설계하고자 한다.

2. MARC의 현황과 문제점

급속히 증가하고 있는 인터넷의 정보자원을 효과적으로 탐색하기 위한 기법의 개발이 진행되면서 새로운 검색모형으로 다양한 분야에서 각종 메타데이터가 제안되고 있다. 가장 오랜 역사를 지닌 대표적인 메타데이터 포맷으로 전통적인 도서관 서지 레코드를 위하여 이미 널리 사용되고 있는 MARC를 들 수 있다. 도서관은 소장 자료의 편목과 분류를 위해 표준화된 시스템을 이용하는데, 온라인 목록을 지원하는 많은 도서관의 협동을 이루는 기반구조가 MARC 레코드이

다.

MARC 포맷은 표준적인 목록 레코드 형식으로 각 나라마다 국가 표준을 제정해 놓고 있다. 최초의 MARC 포맷은 미국의 도서관이 1968년 공표한 LC MARC II이며, 이후 USMARC으로 개정되어 새로운 형태의 정보자료나 통제방식을 지원하도록 확장 개발되어 왔다. 1999년에 USMARC은 캐나다의 CANMARC과 함께 MARC21로 통합되었다[1]. 우리나라에서는 국립중앙도서관에서 1980년에 LC MARC 포맷과 ISO 2709(국제표준화기구)로 제안된 서지정보 교환용 포맷을 토대로 하여 우리나라 문헌의 특성을 수용한 KOR-MARC 즉 "한국문헌자동화목록법 단행본 실험용 포맷"을 개발하여 1993년에 KS로(KS C 5867-1993) 제정하였으며 이후 연속간행물용(1994), 비도서자료용(1996), 전거통제용(1999), 소장정보용(1999), 고서용(2000) 포맷이 각각 KS로 제정되었다.

MARC 레코드의 구조는 네 부분으로 구성되는데 첫째, 고정길이(24문자) 리더(leader)는 레코드 처리를 위한 정보를 제공하는 데이터 요소들인 레코드 길이, 레코드 상태, 레코드 형태, 서지수준, 입력수준, 목록기술형식 등을 포함한다. 둘째, 디렉토리는 어떤 필드가 어느 위치에 나타나며 길이가 얼마인지를 지시해 주는 데이터가 기입된다. 셋째, 제어필드는 레코드를 제어하는 데이터들을 기술한다. 넷째, 데이터필드는 지시기호, 식별기호, 데이터요소가 기술된다[2].

MARC 포맷은 현재 66개 국가에서 사용되고 있다. 그러나 MARC 포맷은 도서관 위주의 목록에 기반한 포맷으로서 도서관 이외의 출판사, 정보센터 등의 관련 기관과의 호환성에 약점을 지니고, 서지 레코드간의 연결기능이 부족하며, 국제적인 서지 레코드의 교환의 경우 별도의 변환프로그램을 필요로하게 된다[3]. 또한 메타데이터로서의 MARC은 인터넷 자원을 수용하기 어렵기 때문에 새로운 목록방식과 지원도구를 요구하여 자원의 기술방식은 간략화하되 탐색방식을 지능화하는 지식베이스 기반의 접근도구를 필요로하게 되었다[4].

MARC의 문제점들을 보완할 수 있는 메타데이터로 SGML, XML 등을 들 수 있다. 특히 XML은 구성과 문법이 복잡하여 사용자가 습득하기 어려운 SGML을 간소화한 표준이라고 할 수 있으며 사용자가 문서의 요소와 속성, 개체를 선언할 수 있는 유연한 DTD를 지원한다.

<그림 1>은 MARC 레코드의 예이며, 각 필드에 대한 이해가 용이하도록 정렬한 것이 <그림 2>이다.

이 MARC 레코드의 각 필드를 구조화하여 XML로 표현하면 <그림 3>과 같다.

```
00935cam 22002654a 45000010009000000050017000090080041000269060045000679250044001129
5501410015601000170029702000150C314040001800329042000800347030002
30035508200160037810000220039424500290041626000360044530000370048
144000400051850002000558650005500578650003600633 12297319
20010327090511.0 010202s2001 nyua 001 0 eng a7
bcb corngnew d1 eocnp l20 gy-gencatlg 0 aacquire b2 shelf copies
xpolicy default apc20 to sa00 02-02-01; se45/se02 02-07-01; se05 to
Dewey 02-08-01; aa05 02-08-01; CIPver. sb17 3-26-01 ; to BCCD
03-26-01; sb00 03-27-01 a 2001030152 a0071371885 aDLC
cDLC dDLC apcc 00 aHD30.37 b.S53 2001 00 a2005.7/2 221 1
aSimon, Solomon H. 10 aXML / cSolomon H. Simon. aNew York :
bMcGraw-Hill, c2001. axxviii, 259 p. :bill. ; c24 cm. 0 aEmerging
business technology series aIncludes index. 0 aBusiness enterprises
xComputer
```

<그림 1> MARC 레코드의 예

```
000 00935cam 22002654a 4500
001 12297319
005 20010327090511.0
008 010202s2001 nyua 001 0 eng
006 $a75bcbc$coringnew$d1$eocnp$b20$p$y-gencatlg
025 0 $aacquire$b2 shelf copies$xpolicy default
055 $apc20 to sa00 02-02-01; se45/se02 02-07-01; se05 to
Dewey 02-08-01; aa05 02-08-01; CIPver. sb17 3-26-01 ; to BCCD
03-26-01; sb00 03-27-01
010 $a 2001030152
020 $a0071371885
040 $aDLC$cDLC$bDLC
042 $apcc
050 00 $aHD30.37$b.S53 2001
082 00 $a005.7/2$221
100 1 $aSimon, Solomon H.
245 10 $aXML /cSolomon H. Simon.
260 $aNew York :$bMcGraw-Hill,$c2001.
300 $axxviii, 259 p. :bill. ;c24 cm.
440 0 $aEmerging business technology series
500 $aIncludes index.
650 0 $aBusiness enterprises$xComputer networks$vSoftware.
650 0 $aXML (Document markup language).
```

<그림 2> 그림1을 ASCII로 변환 · 정렬

```
<?xml version="1.0" encoding="UTF-8" standalone="yes" ?>
<marc>
  <record>
    ...
      <field type="100" i1="1" i2="">
        <subfield type="a">Simon, Solomon H.</subfield>
      </field>
      <field type="245" i1="1" i2="0">
        <subfield type="a">XML</subfield>
        <subfield type="c">Solomon H. Simon.</subfield>
      </field>
      <field type="260" i1="" i2="">
        <subfield type="a">New York :</subfield>
        <subfield type="b">McGraw-Hill.</subfield>
        <subfield type="c">2001.</subfield>
      </field>
      <field type="300" i1="" i2="">
        <subfield type="a">xxviii, 259 p. :</subfield>
        <subfield type="b">ill. ;</subfield>
        <subfield type="c">24 cm.</subfield>
      </field>
    ...
  </record>
</marc>
```

<그림 3> 그림1의 XML 변환 예

3. XML

XML(eXtensible Markup Language)은 인터넷에서 주로 사용되는 HTML의 한계를 극복하고 기존의 SGML이 갖는 복잡함을 해결하기 위하여 웹 상에서 구

조화된 문서를 전송 가능하도록 설계된 표준화된 멀티미디어 전자 문서 형식으로 1996년 W3C(World Wide Web Consortium)에서 처음 제안되었다.

XML은 웹 상에서 데이터 교환을 위해 제안된 표준언어로서 DTD를 통하여 문서 자체에 문서의 구조를 기술한다. 문서의 구조를 사용자가 원하는대로 정의할 수 있으므로 이러한 구조의 유동성은 모든 형태의 데이터가 XML로 기술될 수 있도록 한다. 따라서 XML은 인터넷 상의 모든 데이터가 동일한 형태로 통합, 저장, 처리될 수 있는 기반을 제공한다. 요컨대, XML은 구조화된 정보를 교환하는데 용이할 뿐만 아니라 범용적인 시스템으로 작업을 제공하는 효과적인 수단이 된다고 할 수 있다.

XML의 정교한 구문법은 다른 포맷으로의 쉬운 변환을 가능하게 하므로 XML은 과거와 현재의 포맷을 연결하는 중요한 역할을 한다. 또한 XML은 MARC 데이터베이스를 더블린 코어나 GILS와 같은 포맷으로의 변환을 용이하게 할 수 있다[5].

4. KORMARC의 XML DTD 설계

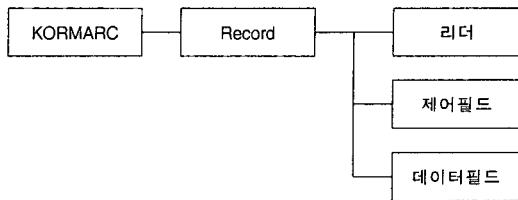
MARC 포맷을 XML로 변환하면 기존의 서지 레코드를 웹 환경으로 쉽게 이식할 수 있을뿐만 아니라 시스템 간의 호환성을 확보할 수 있으며, 메타데이터에 대한 새로운 접근환경을 제공할 수 있게 된다. MARC의 서지 레코드는 논리구조에 따라 기술되고 또한 여러 DTD에서 서지요소의 기술구조를 제시해주고 있으므로 XML을 MARC에 적용하는 것은 큰 어려움이 없다고 볼 수 있다. 다만 DTD를 이용한 MARC 포맷과 XML 포맷간의 변환 과정에서 데이터의 손실이 없어야 하며 현재 MARC에서 적용한 표준을 모두 수용할 수 있어야 한다[4].

DTD를 정의하기 위해서는 먼저 KORMARC 레코드의 구조를 파악할 필요가 있다. KORMARC 레코드의 최상위 구조를 도시하면 <그림 4>와 같다.



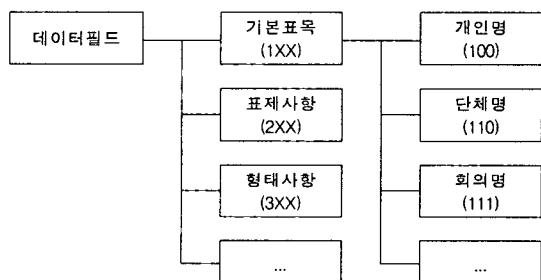
<그림 4> KOBMABC 레코드 구조

이 중에서 디렉토리 부분은 변환프로그램에서 자동으로 계산될 수 있으므로 DTD에서 제외하고, 이를 계층화한 KORMARC의 DTD 구조도는 <그림 5>와 같다.



<그림 5> KORMARC의 구조 계층화

데이터필드는 각각의 세부 필드로 전개되므로 이를 구조화하여 각 필드의 특성을 기준으로 기본표목(1XX), 표제 및 표제관련사항(2XX), 형태사(3XX), 총서사항(4XX), 주기사항(5XX), 주제명부출표목(6XX), 주제명, 총서명 이외의 부출표목, 연관기입필드(7XX), 총서부출표목(8XX) 등으로 구분하며, 각 필드는 다시 하위필드로 세분되어 Tree 구조를 갖게 된다. 예컨대, 기본표목(1XX)은 개인명(100), 단체명(110), 회의명(111), 통일서명(130)으로, 표제 및 표제관련사항(2XX)는 통일서명(240), 서명저자사항(245), 판사항(250), 수치데이터사항(255), 발행사항(260), 발행 예정일자(263) 등의 하위필드로 계층화된다. <그림 6>은 KORMARC 레코드의 데이터필드 중 기본표목(1XX) 필드 및 하위필드의 구조를 계층화한 것이다. DTD에서 하위필드도 독립적인 엘리먼트로 정의함으로써 KORMARC의 의미 구조를 쉽게 파악할 수 있다.



<그림 6> 기본표물(1XX) 펌드 구조의 계층화

한편 KORMARC에서 명시하고 있는 각 필드의 적 용수준(필수, 해당시 필수, 재량)과 반복사용여부(반복, 반복불가)는 XML DTD의 엘리먼트 내용 모델에서 제공하는 출현빈도를 적용한다. 리더와 제어필드는 반드시 출현하고 그 출현 횟수도 1회로 규정하였으며 데이터필드의 경우에는 ‘필수’로 사용되는 하위요소가 하나라도 포함되어 있을 때에만 없음(반드시 1회 출현)으로 하고, 그 외에는 <표 1>과 같이 '?'(출현하지 않거나 1회 출현) '+'(1회 이상 출현) '*'(출현하지 않거나 1회 이상 출현)

거나 1회 이상 출현) 등의 출현빈도를 갖게 한다.

<표 1> KORMARC 필드의 적용수준/반복사용여부

적용수준	반복사용여부	엘리먼트 빈도 지시자
필수	반복불가	없음
해당시필수	반복불가	?
재량	반복불가	?
필수	반복	+
해당시필수	반복	*
재량	반복	*

본 논문에서는 이러한 규칙들을 바탕으로 KORMARC의 XML DTD를 정의하였다. <그림 7>은 KORMARC의 기본표목용 1XX 필드 중 개인명 기본표목인 100 필드와 단체명 기본표목인 110 필드의 DTD 정의 부분을 나타낸 것이다.

```

<!ELEMENT kormarc-main-entry (kormarc100? , kormarc110? , kormarc111? , kormarc130?)>
...
<!ELEMENT kormarc100 ((kormarc100-a | kormarc100-b | kormarc100-c | kormarc100-d | kormarc100-f | kormarc100-g | kormarc100-i)*)>
<!ATTLIST kormarc100 name CDATA #FIXED "기본표목--개인명">
<!ATTLIST kormarc100 i1 (i1-0 | i1-1 | i1-2 | i1-3 | i1-4) #REQUIRED>
<!ATTLIST kormarc100 i2 (i2-0 | i2-1 | i2-blank) #REQUIRED>
<!ELEMENT kormarc100-a (#PCDATA)>
<!ATTLIST kormarc100-a name CDATA #FIXED "개인명">
<!ELEMENT kormarc100-b (#PCDATA)>
<!ATTLIST kormarc100-b name CDATA #FIXED "이름에 포함되어 세계(세계)를 자칭하는 숫자">
<!ELEMENT kormarc100-c (#PCDATA)>
<!ATTLIST kormarc100-c name CDATA #FIXED "이름과 관련된 칭호 및 기타 명칭">
<!ELEMENT kormarc100-d (#PCDATA)>
<!ATTLIST kormarc100-d name CDATA #FIXED "생몰년">
<!ELEMENT kormarc100-e (#PCDATA)>
<!ATTLIST kormarc100-e name CDATA #FIXED "역할이">
<!ELEMENT kormarc100-f (#PCDATA)>
<!ATTLIST kormarc100-f name CDATA #FIXED "역조">
<!ELEMENT kormarc100-g (#PCDATA)>
<!ATTLIST kormarc100-g name CDATA #FIXED "한국 및 중국의 세계(세계)">
<!ELEMENT kormarc100-i (#PCDATA)>
<!ATTLIST kormarc100-i name CDATA #FIXED "언어">
<!ELEMENT mrcb110 ((mrcb110-a | mrcb110-b | mrcb110-c | mrcb110-d | mrcb110-e | mrcb110-f | mrcb110-g | mrcb110-k | mrcb110-l | mrcb110-n | mrcb110-p)*)>
<!ATTLIST mrcb110 name CDATA #FIXED "기본 표목--단체명">
<!ATTLIST mrcb110 i1 (i1-0 | i1-1 | i1-blank) #REQUIRED>
<!ATTLIST mrcb110 i2 (i2-0 | i2-1 | i2-blank) #REQUIRED>
<!ELEMENT mrcb110-a (#PCDATA)>
<!ATTLIST mrcb110-a name CDATA #FIXED "기본요소">
<!ELEMENT mrcb110-b (#PCDATA)>
<!ATTLIST mrcb110-b name CDATA #FIXED "하위기관">
<!ELEMENT mrcb110-c (#PCDATA)>
<!ATTLIST mrcb110-c name CDATA #FIXED "회의개최지">
<!ELEMENT mrcb110-d (#PCDATA)>
<!ATTLIST mrcb110-d name CDATA #FIXED "회의일자나 조약의 서명일자">
<!ELEMENT mrcb110-e (#PCDATA)>
<!ATTLIST mrcb110-e name CDATA #FIXED "역 할이">
<!ELEMENT mrcb110-g (#PCDATA)>
<!ATTLIST mrcb110-g name CDATA #FIXED "기타정보">
<!ELEMENT mrcb110-k (#PCDATA)>
<!ATTLIST mrcb110-k name CDATA #FIXED "형식부표목">
<!ELEMENT mrcb110-l (#PCDATA)>
<!ATTLIST mrcb110-l name CDATA #FIXED "언어">
<!ELEMENT mrcb110-n (#PCDATA)>
<!ATTLIST mrcb110-n name CDATA #FIXED "권차 또는 회차">
<!ELEMENT mrcb110-p (#PCDATA)>
<!ATTLIST mrcb110-p name CDATA #FIXED "권차서명 또는 회차서명">
...

```

<그림 7> KORMARC 기본표목필드(1XX)의 DTD 정의

5. 결론

MARC을 XML로 변환하는 데 따르는 기대효과로서는 첫째, 서지 레코드의 작성 및 다른 포맷으로의 생성이 가능하고 둘째, 웹브라우저, 검색엔진, 그리고 잠재적으로 다른 변환이 필요없는 도서관 시스템에 의해 서지 레코드를 표현할 수 있으며 셋째, 데이터 손실없이 XML과 MARC 상호간의 변환이 가능하고 넷째, 아시아 문자의 표현, 로마나이즈 및 전거통제 등의 MARC의 많은 문제점들이 해소될 수 있으며[6] 다섯째, 도서관에서 XML이 EDI(Electronic Data Interchange) 표준을 대신할 수 있으며 여섯째, 다른 데이터 소스도 XML구문을 사용함으로써 통합과 처리를 더 쉽게 할 수 있다고 하겠다[7].

MARC은 자기테이프가 저장매체로 주로 사용되던 시대에 개발된 포맷이어서 현재에 이르러 많은 문제점을 수반하게 되었다. 즉, 데이터 교환 및 저장매체 기술의 변화를 거의 반영하지 못하므로 새롭게 요구되는 내용을 표현하는 것이 불가능하기 때문이다. 본 논문에서 설계한 KORMARC의 XML DTD를 기반으로 KORMARC 레코드를 XML로 변환할 수 있는 프로그램의 개발 및 구현이 향후 연구과제로 요구된다.

[참고문헌]

- [1] Library of Congress, MARC SGML and XML, <http://lcweb.loc.gov/marc/marcsgml.html>
- [2] 국립중앙도서관, 한국문학자동화목록형식 및 기술 규칙, <http://www.nl.go.kr/kormarc/kormarc.html>
- [3] 조윤희, “XML 기반 디지털도서관 구현에 관한 연구”, 「제7회 한국정보관리학회 학술대회논문집」 pp.79-82, 2000.
- [4] 문헌정보처리연구회 편, 메타데이터의 형식과 구조, 문헌정보처리연구회, 1998.
- [5] John Robert Gardner, eXploring What's neXt : XML, Information Sciences, and Markup Technology, http://vedavid.org/xml/docs/eXploring_xmlandlibraries.html
- [6] K.T. Lam, Moving From MARC to XML, <http://ihome.ust.hk/~lblkt/xml/marc2xml.html>
- [7] Dick Miller, XML: libraries' strategic opportunity, <http://www.ljdigital.com/xml.asp>