MARC과 SGML의 통합에 대한 연구

Incorporation of MARC with SGML

김 태 수 (연세대 문헌정보학과)
최 석 두 (이화여대 문헌정보학과)

Tae Soo Kim (Dept. of Lib. & Inf. Sci., Yonsei Univ.)
Suk Doo Choi (Dept. of Lib. & Inf. Sci., Ewha Womans Univ.)

정보처리기술이 발전하고 이용자의 정보요구가 다양해짐에 따라 MARC의 형식에 한계를 느끼고 全文과 구조적 데이터기술시 많은 장점을 갖고 있는 SGML에 대한 선호도가 높아지고 있다. SGML형식의 이점이 많다고 바로 모든 데이터와 시스템을 SGML형으로 변환할 수는 없다. 그러나 MARC은 도서관분야에서 지대한 역할을 하였지만 일정기간동안 MARC과 SGML이 공존할 뿐 궁극적으로는 SGML형식으로 바뀌게 될 것이다. 본고에서는 MARC의 문제점, SGML로의 이행과정, MARC과 SGML의 융합관계 등을 중심으로 논한다.

1 서론

원래 "마크업"이란 편집자와 인쇄디자이너가 텍스트의 레이아웃이나 서체에 대한 지시사항의 기입을 말하는 것으로 내용이 아니라 물리적인 모양에 관한 모든 것을 지시하는 것이다. 이와 같은 기능은 "절차적 마크업"(procedural markup)이라 하며 최근의 전자출판이나 문서편집기가 거의 갖고 있는 기능이며 대부분 자체의 코드를 갖기 때문에 해당 시스템에서만 통용되고 있다. 한편, 페이지나 스크린의 물리적인 모양과 구별하여 장, 절, 주기, 서지 등 문헌의 구조와 내용을 중심으로 기술하는 "기술적 마크업"(descriptive markup)이 있다.

기술적 마크업언어인 SGML은 1970년에 GML로 처음 개발되어 1986년 10월에 국제표준(ISO 8879)으로 제정되었으며, 다른 마크업언어를 정의하는 틀 혹은 규칙이므로 하나의 메타언어로 취급되고 있다. SGML에서는 문헌에서 필요한 요소와 실제로 출현하는 요소에 대한 구조화 정보를 "문헌유형정의"(DTD)를 이용하여 처리과정과 독립적으로 기술할 수 있으며, 이 기술의 정당성은 구문분석기로 확인할 수 있다. 처리과정과 독립적이므로 DTD만 알면 SGML로 코드화된 데이터를 상이한 하드웨어와 소프트웨어상의 여러 가지 응용시스템에서 다양한 방법으로 이용할 수 있다.

데이터의 기술언어로서 MARC 대신에 이와 같은 장점을 가지고 있는 SGML을 사용하려는 경향은 정보처리기술이 발전하고 이용자의 정보요구가 다양화됨과 아울러 이들의 정보요구에 부응하는 정보처리를 위해서는 MARC 형식이 한계가 있기 때문이다. 본고에서는 MARC의 문제점, SGML로의 이행과정, MARC과 SGML의 융합관계 등을 중심으로 논하고자 한다.

2 SGML경향과 MARC의 문제점

많은 도서관과 연구프로젝트에서 MARC의 문제점을 해결하기 위하여 신규 및 기존데이터를 SGML형식으로 변환하기 위한 연구를 수행하고 있다. TEI프로젝트, 버클리의 Finding Aids프로젝트 및 컬럼비아대학의 DIAP프로젝트를 예로 든다.

텍스트코드화 프로젝트(The Text Encoding Initiative: TEI)(http://www-tei.uic.edu/orgs/tei/)는 전자텍스트의 다양성을 줄이고 기계처리를 단순화하며 정보자원을 쉽게 공유할 수 있는 전자텍스트의 작성과 교환지침을 만들기 위해 미국, 유럽, 캐나다 등의 정부기관, 학협회, 재단 등이 참가하는 국제적인 협력사업으로 1987년 11월에 Poughkeepsie의 Vassar대학에서 시작되었다. 이 회의에서 신규문헌과 기존문헌의 전자텍스트 코드화 및 교환지침을 위한 공통코드화계획에 합의하고 "Poughkeepsie원칙"을

발표하였으며, 1988년부터 연구를 시작하여 1990년 6월에 초안을 출간하였다. 즉시 개정작업에 착수하여 1992년 4월부터 1993년 11월에 걸쳐 지침의 개정안이 장별로 발표되었으며, 1994년 5월 지침의 공식판이 결정되었다. 전자화된 텍스트는 자연언어처리, 정보검색, 하이퍼텍스트, 전자출판, 다양한 문헌 및 사적 분석, 사전편찬 등 응용분야가 점점 넓어지고 있으며, 이와 같은 응용분야 뿐만 아니라 앞으로 생길 수 있는 어떤 목적에도 사용될 수 있도록 하는 것이 TEI프로젝트의 목표이다.

지금까지 개발된 TEI의 지침을 보면 TEI문헌은 TEI헤더부분과 DTD에 따라 코드화된 텍스트본문으로 구성된다. TEI헤더는 파일기술부(전자파일의 완전한 서지기술), 코드화기술부(원자료와 코드화된 텍스트와의 관계), 프로파일기술부(텍스트의 비서지적 측면에 대한 상세한 기술), 개정내역기술부(파일에 대한 모든 변경내역의 요약)의 4부분을 가지며, 이 헤더부분은 MARC레코드와 밀접한 관련을 갖는다. 이용되는 태그수는 약 400여개, 헤더에 이용되는 것은 약 60여개, 기본태그는 약 100여개 정도이다. 텍스트(全文)의 구조에 대해서는 ISO/DIS 12083-1994를 참조하고 있다.

TEI는 순수한 서지정보만을 보면 TEI헤더는 임시목록레코드와 매우 유사하다. 즉, TEI헤더의 설계자는 헤더의 정보로 임시MARC목록레코드를 생성할 수 있도록 하였다. MARC레코드와 TEI헤더의 가장 중요한 차이는 각각의 기능에 있다. 도서관계의 요구와 노력에도 불구하고 MARC레코드는 기본적으로 목록카드의 전자판이다. 목록카드는 복잡한 서지데이터를 포함하고 있는 물리적인 대상에 대한 단위레코드이므로 목록카드는 물리적인 대상을 가리킨다. TEI헤더는 완전한 서지정보 뿐만 아니라 비서지정보도 자동이나 수동으로 분석할 수 있다. TEI에서 프로파일, 코드화, 개정내역기술부에 기술되는 이들 서지정보가 MARC레코드에서는 비구조화 주기필드(5xx)에 기술된다. 일반적으로 주기필드는 구조화되어 있지 않아 기계적인 검색과 분석에 도움이 되지 않으나 적정한 형식으로 각 기술부에 기술됨으로써 기계에 의한 분석과 처리가 가능하게 된다. 따라서 필요할지가 의문이지만 필요하다면 TEI헤더의 정보를 분석하여 자동으로 완전한 목록을 만드는 지능형 에이전트를 만들 수 있을 것이다.

"Finding Aids for Archival Collections"(http://sunsite.berkeley.edu/FindingAids/)는 "The Berkeley Finding Aid Project: BFAP"라는 이름으로 1993년 가을부터 시작한 연구이며, 보존문서관, 박물관, 도서관의 코드화표준을 개발하기 위한 공동프로젝트이다. Finding Aids란 관련자료를 기술하고, 통제하며, 액세스에 사용되는 문헌을 말한다. 집서수준의 정보액세스나 항해를 위한 계층구조에서 Finding Aids는 서지레코드와 1차자료의 중간에 위치하게 된다. 서지레코드는 Finding Aids로 안내하고, Finding Aids는 1차자료로 안내한다. 이 프로젝트의 목표는, 첫째, SGML DTD의 형식으로 Finding Aids의 전형적인 코드화 표준을 만드는 일이며, 둘째, Finding Aids의 전형적인 데이터베이스를 구축하는 일이다. SGML과 FINDAID DTD를 사용하던 것을 1995년 후반에는 Encoded Archival Description(EAD)으로 개칭하고 EAD DTD의 개발을 완료하고 시험을 거쳐 배포하고 있다.

MARC은 원래 개별 서지에 적용되는 기술사항과 액세스정보를 찾기 위해 설계된 것이므로, MARC구조에서는 하향서지구조로 복잡한 집서에 대한 액세스를 기술하게 되면 즉시 과중한 부담을 안게 된다. 적어도 제2 서지수준은 가능하겠지만 그런 정보는 한계가 있다. 이 문제를 해결하는 방법은 다양한 서지수준에서 다중 링크를 각 레코드에 적용하는 것이지만 시스템내 및 시스템간의 제어가 매우 어렵다. Finding Aids는 계층적으로 구조화된 문헌이며, 평면구조인 MARC은 부적당하다고 판단되어, 레코드 코드화에 SGML이 채택되었다.

Columbia University Digital Image Access Project (DIAP)(http://www.cc.columbia.edu/cu/libraries/indiv/avery/diap.html)는 컬럼비아대학도서관이 개발하고 있으며 디지털이미지에 대한 서지데이터를 축적하고 액세스하기 위한 새로운 모형이다. DIAP팀은 요약서지정보 뿐만 아니라 필요하다면 상세한 계층 및 版別 관련데이터를 SGML로 코드화된 서지(메타데이터)레코드로 축적할 수 있다는 것을 제안하였으며, 이 레코드에는 실제적인 디지털문헌, 다른 관련 서지레코드, 심지어는 외부의 전자출판물, 데이터베이스, 수치파일 등과 같은 관련 디지털자료에 대한 링크를 포함시킬 수 있다고 하였다. 이 새로운 형식의 레코드를 위한 SGML Catalog Record(SCR)가 구체적으로 제안되었다. SCR은 보다 융통성있게 데이터요소 클러스터들을 통합함으로써 계층관계요소를 분할하고 판별정보를 분리하며 개별레코드로 만들어야 하는 현재의 AACR2/USMARC모형보다 복잡한 서지정보를 기술하고 표현하는 데 적합하다고 판단하고 있다.

3 MARC/SGML 프로젝트

지금까지 도서관목록에서의 검색내용이란 저자, 서명, 주제, 주요어의 불탐색 등이 주였다. 그러나 최근에는 Web에 Alta Vista, Excite, Lycos, InforSeek, Inktomi, Aliweb, Harvest, Magellan, Open Text, Web Crawler, WWW Worm, Yahoo 등의 수많은 색인서버들이 나타났으며, 완전한 것은 아니지만 많은 시스템들이 가중치탐색, 단어의 위치나 빈도를 이용한 적합성 피드백, 자동어미절단, "more of the same" 엘고리듬, 자연언어 질문시스템, 개념기반탐색, 의미트리 등의 새로운 기법들이 사용되고 있다. 현재 도서관이 가지고 있는 데이터베이스를 인터넷, Web 혹은 관련시스템으로 변경하려고 할 때 전통적인 도서관시스템용으로 설계된 MARC은 여러 부분이 경직되어 있으며, 유일한 출구가 ANSI/NISO Z39.50-1995로 너

무나 좁다. 따라서 MARC의 SGML판을 만들 필요가 생기게 되었다.

MARC/SGML프로젝트는 미국국회도서관의 자문위원회(이하 "자문위원회"라 한다)가 중심이 되어 1990년부터 MARC레코드용 SGML DTD를 개발하기 시작하였으며, 개인, 기업, 기관이 참가하고 있다. 공식적으로 채택된 것은 아니지만 자문위원회에서 정한 설계원칙은 다음과 같다(Davis, 1996).

- MARC DTD는 실제 MARC에서 SGML로, SGML에서 MARC으로 정보의 손실 없이 완전하게 양방향 변환이 가능해야 한다.
- MARC DTD는 현 MARC표준을 따라야 하며 유지보수도 MARC표준과 병행되어야 한다.
- MARC레코드의 구조 요소(예를 들면, 레코드의 길이, 디렉토리)는 MARC/SGML레코드를 MARC레코드로 변환할 때 다시 계산할 수 있으므로 MARC/SGML이 가질 필요는 없다.
- 상기 변환프로그램은 본 프로젝트에서 개발한다.
- MARC/SGML레코드는 SGML문헌 내에 포함시킬 수도 있으며 메타데이터로서 독립시킬 수도 있다.

그러나 실제 상황에서 다음과 같은 몇 가지 문제점이 나타나고 있다. 첫째, MARC은 이산적인 요소들로 구성되어 있으며 대부분 순서에 독립적인 반면, SGML은 상대적으로 계층적이며 데이터요소의 정의는 문맥과 밀접한 관계를 가지고 있다. 둘째, MARC은 실제적으로 쓰이지 않는 데이터요소를 많이 가지고 있다. 이것을 모두 MARC/SGML표준에서 수렴해야 할 것인지, 현재 유효한 것만 지원해야 할 것인지 정하기가 어렵다. 셋째, MARC은 로컬정보를 명확하게 지원하고 있다. MARC/SGML에서는 이들 요소를 어떻게 처리할 것인가? 넷째, MARC/SGML은 MARC표준에만 국한할 것인가 아니면 다른 것을 추가할 것인가? 끝으로, MARC에서 Web을 위해 개발된 필드 856의 URL, URN, PURL과 같은 링크를 SGML에서도 지원해야 할 것인가?

자문위원회에서는 제기된 문제들을 고려하여 MARC DTD를 개발하여 제안하였다. 이미 USMARC.DTD와 변환프로그램이 Jerome McDonough 등에 의해 개발되었으며 MARC레코드와 SGML형식간에 정보의 손실을 최소화하면서 양방향으로 자동변환할 수 있도록 설계되었다. USMARC레코드를 SGML형식으로 바꾸는 프로그램으로 marc2sgml, 그 역방향 변환프로그램으로 sgml2marc이 개발되어 사용되고 있으나 외국어로 된 레코드는 제외되고 있다. USMARC.DTD는 익명ftp인 "ftp://library.berkeley.edu/pub/sgml/marcdtd/usmarc.dtd"에서 얻을 수 있다. 서명과 관련되는 부분을 발췌해보면 다음과 같다.

```
<!DOCTYPE USMARC [
<!ELEMENT USMARC - - (Leader, Directory, VarFlds)>
<!ELEMENT USMARC  Material (BK|AM|CF|MP|MU|VM|SE) "BK"
                  id      CDATA ' #IMPLIED>
<!ELEMENT Leader - O (LRL, RecStat, RecType, BibLevel, ...)>
<!ELEMENT Directory - O (#PCDATA)>
<!ELEMENT VarFlds - O (VarCFlds, VarDFlds)>
<!ELEMENT VarDFlds - O (NumbCode, ..., Titles, ....)>
<!ELEMENT Titles - O  (Fld210?, ..., Fld242*, Fld243?, Fld245, ...)>
<!ELEMENT Fld245 - O (Six?, (a|b|c|f|g|h|k|....))>
<!ATTLIST  Fld245 AddEnty (No|Yes|Blank)          #IMPLIED
                  NFChars (0|1|2|3|4|5|6|7|8|9|Blnk)  #IMPLIED>
<!ELEMENT a - O   (#PCDATA)>
<!ELEMENT b - O   (#PCDATA)>
<!ELEMENT Six - O (#PCDATA)>  ]>
```

이를 바탕으로 컬럼비아대학이 Web에서 직접사용하는 것을 목표로 1996년초부터 MARC/SGML 시험레코드를 생성하고 있으며, HTML레코드도 포함시킬 계획이다. 자문위원회의 형식과 컬럼비아대학도서관 형식에는 차이가 있다. 첫째, 컬럼비아대학에서는 MARC/SGML형식에서 상이한 版, 어떤 집서의 부분을 기술하고 외부의 디지털오브젝트와 직접 링크시키기 위하여 복수의 서브레코드를 지원할 수 있도록 확장하여 사용하고 있다(다음 예의 Subrec 참조). 둘째, 디렉토리와 레코드의 길이는 MARC/SGML레코드를 MARC레코드로 변환할 때 다시 계산할 수 있으므로 제외하고 있다. 셋째, 필드 및 서브필드태그명을 다르게 사용하고 있다. 컬럼비아대학 USMARC DTD의 예를 MARC의 서명과 관련되는 것만 발췌해보면 다음과 같다(http://www.cc.columbia.edu/cu/libraries/inside/project/sgml/).

```
<!DOCTYPE USMARC [
<!ELEMENT USMARC - - (Leader, VarFlds, SubRec*)>
<!ELEMENT USMARC  Material (BK|AM|CF|MP|MU|VM|SE) "BK"
                  id      CDATA    #IMPLIED>
<!ELEMENT SubRec - - (Leader?, VarFlds?)      >
<!ATTLIST  SubRec TYPE (version|component|CDATA) #REQUIRED
           SubRec LEVEL CDATA #IMPLIED
           SubRec CLASS CDATA #REQUIRED
           SubRec FIELDNAME CDATA "Subrecord">
<!ELEMENT VarFlds - O (VarCFlds, VarDFlds)>
<!ELEMENT VarDFlds - O (NumbCode, ..., Titles, ....)>
<!ELEMENT Titles - O  (mfb210?, ..., mfb242*, mfb243?, mfb245, ...)>
<!ELEMENT mfb245 - O (mbs.Six?, (mbs.a|mbs.b|mbs.c|mbs.f|mbs.g|
                     mbs.h|....))>
<!ATTLIST mfb245 AddEnty (No|Yes|Blank)          #IMPLIED
                 NFChars (0|1|2|3|4|5|6|7|8|9|Blnk)  #IMPLIED>
<!ELEMENT mbs.a - O   (#PCDATA)>
<!ELEMENT mbs.b - O   (#PCDATA)>
<!ELEMENT mbs.Six - O (#PCDATA)>  ]>
```

4 MARC형식의 보완

이용자가 저작, 저작의 다양한 판, 저작의 부분, 저작의 全文 등 다양한 수준으로 접근할 수 있도록 하기 위해서

는 계층적인 접근이 필요하게 되었다. MARC은 분명히 새로운 도서관시스템에서 요구하고 있는 기능을 충족시키지는 못하고 있다. 1990년 이래 HTML은 이미 사서에게 낯설지 않을 것이다. HTML은 SGML의 응용이며, 문헌내의 이미지, 소리, 비디오를 포함할 뿐만 아니라 별개의 문헌, 이미지, 소리파일과 하이퍼링크를 가질 수 있다.

그렇다면 MARC형식은 더 이상 이용가치가 없으며 얼마 가지 않아 없어질 것인가? 그렇지는 않다. 수십억의 MARC레코드가 시스템상에 있으며, 이들을 SGML로 바꾸는 시간과 컴퓨터자원상의 비용은 상상도 못할 것이다. MARC은 가장 효율적으로 이용할 수 있는 방향으로 진화할 것이다. 즉, 도서관시스템에서 서지데이터를 코드화할 수 있는 형식은 이제 MARC이 전부가 아니기 때문이다(Gaynor, 1996).

미국국회도서관은 USMARC을 개정하였으며 그 중 중요한 것은 필드 856(전자적 위치 및 액세스)을 설정했다는 점이다. 해당 정보를 포함하고 있는 전자적 위치나 그 정보를 얻을 수 있는 위치를 식별하기 위한 것이다.

제1지시기호(0 전자우편, 1 FTP, 2 원격로그인, 3 다이얼업, 7 서브필드 $2에서 명시한 방법, 8 기타)는 필드에 있는 데이터 이외의 데이터를 어떤 액세스방법으로 검색할 수 있는지를 정의하는 값을 갖는다. 정의된 방법은 TCP/IP프로토콜이다. 제1지시기호의 값은 어느 서브필드를 사용할 것인지를 결정한다. 예를 들면, 제1지시기호의 값이 1 (FTP)이면, 서브필드 $d(경로), $f(전자적 이름), $c(압축정보), 및 $s(파일크기)를 사용하게 되며, 이들은 제1지시기호가 2 (원격로그인: Telnet) 일 때는 사용할 수 없다.

이 필드에는 화일의 전자적 전송, 전자잡지의 구독신청, 혹은 전자정보자원에 로그온할 수 있는 충분한 정보를 기술할 수 있다. 경우에 따라서는 원격지 호스트가 정보를 가지고 있어서 이용자는 locator table를 액세스할 수 있는 유일한 데이터식별요소만을 기술할 수도 있다. 위치데이터요소가 다양하거나($a, $b, $d) 두 가지 이상의 액세스방법을 사용하고 있을 때는 필드 856을 반복할 수 있다. 또한 단일 자료가 온라인 축적 및 검색을 위해 여러 개의 부분으로 분리되어 있는 경우를 제외하고는 전자화일명(서브필드 $f)이 바뀔 때마다 반복할 수 있다.

따라서 설정된 서브필드 코드 중 중요한 것을 보면, $a 호스트명, $b 액세스번호, $c 압축정보, $d 경로, $f 전자적 이름, $j BPS, $k 패스워드, $l 로그온/로그인, $n 서브필드 $a에 있는 호스트의 위치명, $o 오퍼레이팅시스템, $p 포트, $q 파일포맷, $s 파일크기, $t 터미널이뮬레이션, $u URL, $w 레코드제어번호, $2 액세스방법 등이 있다(gopher://marvel.loc.gov:70/00.listarch/usmarc/96-1.doc).

5 결론

미국국회도서관이 목록카드를 배포한 이래 MARC레코드와 관련 표준은 도서관 분야에 지대한 역할을 해왔지만, 전자정보환경은 도서관의 외부에서 형성되고 있을 뿐만 아니라 MARC을 중심으로 하는 도서관은 더 이상 정보혁명의 최전선이 아니라고 보아야 한다. 그러나 SGML 형식의 이점이 많다고 바로 모든 데이터와 시스템을 SGML형으로 변환할 수는 없다. 이와 같은 상황에서 MARC과 SGML의 이용법은 정책에 따라 여러 가지로 나눌 수 있다.

첫째, MARC과 SGML의 어느 한쪽을 중점적으로 채택할 수 있다. MARC을 채택하는 경우에는, 1) 지금까지의 형식을 그대로 사용하는 방법, 2) USMARC과 같이 필드 856 등을 설정하여 MARC형식을 개정하고, MARC형식 데이터와는 관계없이 SGML형식으로 기술된 서지 및 전문데이터는 링크를 이용하여 참조하는 방법이 있다. SGML을 사용하는 경우에는, 1) MARC을 단순히 MARC/SGML형식으로 바꾸는 방법, 2) MARC/SGML형식으로 바꾸고 이미지를 포함하는 全文데이터를 TEI문헌으로 추가하는 방법이 있다(全文의 구조에 대해서는 ISO/DIS 12083-1994 및 ANSI/NISO Z39.59-1988 참조).

둘째, MARC과 SGML을 독립적으로 사용하는 방법이 있다. 모든 데이터는 MARC형식을 사용하고, 全文데이터로 변환되는 것만 SGML을 사용하는 방법이다. 결국 이 방법은 MARC을 채택하는 경우의 2)와 같아질 것이다.

MARC과 SGML은 상당기간 공존하게 될 것이나 궁극적으로는 SGML형식으로 바뀌게 될 것이다. 이를 위하여 KORMARC/SGML을 위한 KORMARC.DTD와 양방향 변환프로그램의 개발이 시급하다고 사료된다.

참고문헌

ANSI/NISO Z39.50-1995. *Information Retrieval Application Service Definition and Protocol Specification for Open System.*

ANSI/NISO Z39.59-1988. *Electronic Manuscript Preparation and Markup.*

Davis, Stephen Paul(1996). *SGML-MARC: Incorporating Library Cataloging into the TEI Environment.* (http://www.columbia.edu/cu/libraries/inside/projects/sgml/sgmlmarc/davis9603.html#1).

ftp://library.berkeley.edu/pub/sgml/marcdtd/usmarc.dtd

Gaynor, Edward(May 8, 1996). *From MARC to Markup: SGML and Online Library systems.*(http://www.lib.virginia.edu/speccol/scdc/articles/alcts_brief.html)

gopher://marvel.loc.gov:70/00.listarch/usmarc/96-1.doc

http://sunsite.berkeley.edu/FindingAids/

http://www-tei.uic.edu/orgs/tei/

http://www.cc.columbia.edu/cu/libraries/indiv/avery/

http://www.cc.columbia.edu/cu/libraries/inside/project/sgml/

ISO/DIS 12083-1994 *Information and Documentation - Electronic Manuscript Preparation and Markup*