95-c-2

전문데이터베이스와 SGML의 관계

A Study on the Relation of Full-text Database and SGML

윤 소영 (숙명여자대학교 문헌정보학과 대학원)
김 성혁 (숙명여자대학교 문헌정보학과)

So Young Yoon, Sung-Hyuk Kim
Dept. of Library and Information Science, Sookmyung Women's University

본 논문은 문서를 기술하는 마크업 언어인 SGML의 개념과 SGML 응용 현황들을 고찰하고 우리 글로 된 전문데이터베이스 구축시 SGML을 이용하기 위하여 필요한 사항들을 제안하였다.

1. 서 론

디지털도서관에 대한 관심이 증가하면서 전문데이터베이스의 구축과 검색, 교환에 대한 연구가 활발히 진행되고 있다. 전문데이터베이스의 교환과 검색을 용이하게 하려는 목적에서 문서의 구조를 분석하여 문서를 구성하는 요소들이 나타내고 있는 의미를 표현하기 위한 마크업 언어에 대한 연구도 그 중 하나이다. 이미 ISO 8879로 지정된 SGML (Standard Generalized Markup Language)은 시스템 환경에 제한되지 않는 범용 마크업 언어로서 전문데이터베이스의 구축뿐 아니라 검색, 교환에 사용되는 메타언어이다.

본 논문에서는 SGML의 개념, SGML과 HTML, TEI, 디지털도서관 구축에 관한 사항들을 살펴보고 우리글로 된 전문데이터베이스 구축시 SGML의 활용을 위하여 필요한 과제들을 제시하고자 한다.

2. SGML의 개념

ISO는 1986년 SGML(ISO 8879)을 문서를 기술하는 표준언어로 채택하였다. SGML을 직역하면 「표준범용마크업언어」가 되며, 문서의 구성 요소를 기술하고 문서의 내용 자체를 표현할 수 있기 때문에 「문서기술언어」라고도 한다.

SGML로 작성된 문서는 SGML선언(Documentation Declaration), 문서형정의(DTD ; Document Type Definition), 실제문서(Document Instance)로 구성되며, 도표나 그림이 존재하는 경우는 그것을 비트패턴화하여 외부 파일에 저장하고, 필요시 외부 엔티티를 사용하여 이 파일을 불러들인다.

SGML선언에서는 문자집합(character set), 처리에 필요한 기억용량(capacity set), 구체적 구문의 유효범위(concrete syntax scope), 사용되고 있는 구체구문(concrete syntax), 사용되고 있는 특성(feature use), 처리에 필요한 응용고유정보(application-specific information) 등의 마크업 속성을 기술한다.

문서형정의는 실제문서를 SGML 구문을 사용하여 정의·기술한 것으로, 책, 논문 등 문서의 종류에 따라 문서구조의 선언문이 달라진다. 문서형정의는 문서요소선언, 속성정의리스트선언, 엔티티선언으로 구성되는데, 문서의 종류에 따라 문서형정의를 준비하여야 하며 외부에서 참조되기도 한다. 그 예는 다음과 같다.

```
< -- 문서요소선언 -->
<!ELEMENT textbook (front, body, rear)>
< -- 속성정의리스트선언 -- >
<!ATTLIST fig id ID #IMPLIED>
< -- 엔티티선언 -->
<!ENTITY SGML "Standard Generalized Markup Language">
< --공개된 문서종류를 사용하는 선언(manual) -- >
<!DOCTYPE manual PUBLIC "-//Cave Press//DTD Manual//EN">
```

실제문서는 문서형정의에 따라 마크업한 실제문서이다. 문서 내에 SGML 특유의 마크를 삽입한 형식으로 되어 있다.

```
<textbook>
<front>
 <title> 정보관리론 </title>
<author> 사공철, 김태수 </author>
<content> 서론
          1. 문헌의 유형
          2. 서지 및 참고자료
.....
</content>
</front>
<body> <chap> 서론
<p> '커뮤니케이션'과 '정보'는 현시대에 있어서 .....
<p>  대인간의 직접 커뮤니케이션은 .....
</chap>
<chap> 서지 및 참고자료
<p> 참고자료는 정보탐색에 사용되는 .....
.....
</chap>
</body>
</textbook>
```

3. SGML의 응용

3.1 SGML과 HTML

HTML(HyperText Markup Language)은 SGML의 하부집합으로 하이퍼미디어 문서를 작성하고 검색하기 위하여 사용되는 언어이다. WWW(World-Wide Web)에서 하이퍼텍스트의 상호교환을 위하여 HTML로 작성된 문서를 인터네트상에 공개한다. 이용자는 HTML로 작성된 문서를 WWW의 인터페이스 소프트웨어인 Mosaic을 이용하여 검색할 수 있다. HTML이 SGML에 따라 작성되므로 SGML의 구문분석기에 의해 구문분석될 수 있다. 따라서 WWW/Mosaic이 인터네트상에서 널리 이용되더라도 NCSA(National Center for Supercomputing Application)가 언급한 것처럼 HTML은 하이퍼미디어 문서를 작성하기 위하여 SGML을 일부 응용한 것이므로 보다 구조화되고 통합된 환경에서 문서를 작성하기 위해서는 SGML을 사용해야 한다.

최근 들어 인터네트상에서 SGML로 작성된 문서를 검색하기 위한 소프트웨어 및 통신 프로토콜에 관한 연구가 진행되고 있는 것도 HTML로는 인쇄매체의 전문을 마크업하여 활용하는데 한계가 있기 때문일 것이다.

3.2 SGML과 TEI

유럽지역은 이미 1960년대에 인문과학분야에 컴퓨터를 도입하기 시작하여 고전의 텍스트 데이터베이스를 대량으로 집적한 아카이브나 코퍼스를 작성하였다. 그러나 각각 독자의 방식으로 데이터베이스화하고 있어서 공통의 코딩스키마(coding scheme)가 없었다. 따라서 공개·공유를 목적으로 한 텍스트데이터베이스의 표준적인 코딩스키마를 작성하게 되었으며 이는 코퍼스 작성의 표준화와도 보조를 맞추게 되었다.

1987년 "텍스트 인코딩 가이드라인(Text Encoding Guidelines)" 이라는 포키푸시회의를 계기로 NEH(US National Endowment for the Humanities)가 기금을 마련하고, ACL(Association for Computational Linguistics), ACH(Association for Computers and the Humanities), ALLC(Association for Literary and Linguistic Computing)가 공동으로 지원하는 국제프로젝트 TEI가 결성되었다. TEI의 구체적인 목적은 다음과 같다.

① 텍스트 데이터베이스의 공통교환 포맷의 작성 및 인코딩에서의 문제점 해결.
② 텍스트를 인코딩할 때, 어떤 요소가 코딩되어야 하며, 또 그것을 어떻게 표현하는지에 관해서 구체적인 조언을 주는 가이드라인을 작성.
③ 주된 인코딩방법을 조사하여 문서를 작성하는 동시에 메타언어를 개발.

TEI는 그들의 구체적인 목적에 부합하는 메타언어로서 SGML(ISO 8879)을 채택하여, 이를 TEI 가이드라인(TEI Guidelines)의 기준언어로 사용하였다. TEI가 SGML을 선택한 이유는 이미 국제표준이 되었다는 장점 외에, 문서형정의를 가지는 기술적인 마크업언어이며, 작성된 문서는 그 자체를 하드웨어나 소프트웨어에 구애됨이 없이 독립적으로 교환을 할 수 있다는 특성 때문이었다.

TEI 문서는 TEI헤더부분과 DTD에 따라 인코드된 텍스트 본문으로 구성된다. 이용되는 태그의 수는 약 400여개, 헤더에서 이용되는 태그는 약 60개 정도이며, 핵심이 되는 기본태그는 약 100개정도이고, 필요한 것을 선택하여 추가할 수 있다. TEI DTD는 메인 DTD(main DTD)와 보조 DTD(auxiliary DTD)로 구성되는데 메인 DTD는 핵심, 기본, 부가 태그 셋(core, base and additional tag sets)으로 구성되며, 보조 DTD는 independent header, writing system declaration, feature system declaration, tag set declaration 등으로 구성된다. TEI 문서의 기본적인 텍스트 구조는 <text>, <front>, <body>, <group>, <back>이며 <group>은 텍스트들을 그룹으로 모아 주기 위하여 사용하는데, 예를 들어, 한 작가가 쓴 여러 편의 수필이 한 문서에 동시에 나타날 때 사용한다.

TEI는 이러한 많은 태그를 일관되게 사용할 수 있도록 산문, 운문, 드라마, 연설문, 인쇄된 사전, 용어데이터베이스의 기본태그셋(base tag set)과 한 문서에 나타나는 여러 종류의 문서형을 처리하기 위한 일반태그셋(generic-base tag set)과 혼합태크셋(mixed-base tag set) 등 8 종류의 기본적인 DTD를 제공하고 있다.

이외에도 SGML을 응용한 신문기사전문 마크업 언어인 UTF(Universal Text Format)가 있으며, 문서의 논리적 구조에 따라 하이퍼텍스트, 멀티미디어, 하이퍼미디어 등을 처리하기 위한 HyTime(Hypermedia/Time-based Structuring Language)이 ISO 표준(ISO 10744:1991)으로 제정되어 있다. 또한 SGML로 마크업된 전문데이터베이스를 인쇄하는데 필요한 포맷을 표준화한 SPDL(Standard Page Description Language), DSSSL (Document Style Semantics and Specification Language)등도 ISO 표준으로 제정되어 있다.

우리 나라도 전문데이터베이스 작성을 위하여 SGML을 사용하는 것과 동시에 TEI와 같

이 텍스트의 특성에 맞는 DTD를 설정하는 것도 필요하다. 이를 위해서 먼저 우리글 텍스트의 구조를 파악하는 일이 선행되어야 할 것이다.

4. 디지털도서관 구축과 SGML

디지털도서관은 컴퓨터와 통신망을 바탕으로 이루어지며, 현재 도서관이 제공하는 있는 서비스보다 많은 정보를 신속하고 정확하게 제공받을 수 있어야 한다. 이러한 디지털도서관을 구성하기 위하여 필요한 요소는 여러 가지가 있겠지만 무엇보다 디지털도서관의 정보원이 되는 전문데이터베이스의 구축을 들 수 있다. 전문데이터베이스를 구축하기 위하여 지금까지는 ASCII 코드로 작성하거나 이미지 처리 등을 사용하였다. 그러나 ASCII와 같은 코드체계를 사용할 경우 컴퓨터 시스템간에 호환이 되지 않는 장애가 발생하였으며, 이미지 처리를 사용할 경우 화상처리는 가능할지라도 문자처리를 위해서는 변환(conversion)이 필요하며 문서의 구조는 파악되지 않는 등의 제한점이 있었다. 따라서 이러한 문제를 해결하기 위하여 문서의 구조를 파악하여 기술하여 주는 마크업언어를 사용하게 되었다. 전문데이터베이스 구축시 여러 마크업언어들중에서 표준으로 제정된 마크업언어인 SGML을 사용하게 되면, DTD를 통해 문서의 구조를 파악하여 문서의 구성요소들이 나타내고 있는 의미까지 파악해 주므로 구축 및 검색이 용이하고, 이미 국제표준으로 채택되어 있어 정보의 공유와 전달이 가능하다. 또한 미국출판자협회(AAP ; Association of American Publisher)가 제공하는 수식/도표에 관한 DTD나 TEI가 제공하는 문서 종류에 따른 DTD를 사용할 수 있는 장점도 가지게 된다. 그러므로 디지털도서관의 전문데이터베이스 구축시 SGML을 사용하는 것이 이러한 여러 이점을 얻을 수 있으므로 바람직할 것이다.

5. 연구과제 및 결론

디지털도서관의 전문데이터베이스의 작성을 위하여 이용되는 표준범용마크업언어인 SGML은 이미 유럽이나 미국 등에서는 실용화되고 있으며, SGML 문서 작성을 위한 SGML 에디터나 SGML 파서와 같은 소프트웨어도 다양하게 개발되어 있다. 우리 나라에서도 SGML에 관련된 연구가 몇몇에 의해 이루어지고 있기는 하지만 아직은 문서의 구조를 파악한다든가 하는 본질적인 문제까지 접근하지는 못한 실정이다.

우리 나라도 디지털도서관이 구축되고 있는 이 시점에서 전문데이터베이스 작성을 위하여 국제표준으로 제정되어 있는 SGML을 빠른 시일 내에 우리나라에 맞게 한글마크업언어화하여 표준으로 제정하여야 한다. 또한 SGML 문서 작성시 실제 업무를 위한 도구로 한글을 처리할 수 있는 SGML에디터, SGML파서, 포맷터 등이 개발되어야 하며, SGML로 작성된 문서를 검색하기 위하여 필요한 데이터베이스 모델도 고려되어야 한다. 그리고 SGML로 작성된 전문데이터베이스는 대용량을 차지하게 되므로 이의 빠른 검색을 위한 색인 및 검색도 고려되어야 한다. 무엇보다 우리글로 된 텍스트 구조의 특성을 파악하여 DTD를 설정해야 할 것이다.

참고문헌

根岸正光, 石塚英弘. 1994. *SGMLの活用*. 東京 : オーム社

ISO 8879:1986. Information Processing Text and office systems - Standard Generalized Markup Language

Sperberg-McQueen, C. M. & Lou Burnard. 1994. *Guidelines for Electronic text encoding and Interchange*. TEI P3. Chicago : TEI.