연재 순서

1. 메타데이터의 개요
2. DC(Dublin Core)
3. GILS(Govenment Information Locator Service
4. IAFA Templates
5. MARC
6. PICS(Platform for Internet Content Selection)
7. RFC 1807
8. SOIF
9. TEI header ····· 이번호
10. URC(Uniform Resource Characteristics
11. Warwick Framework RDF (Resource Description Framework)
12. 메타데이터 향후 방향

전자문헌의 작성과 교환을 위한 메타 데이터

TEI는 학술 연구에 이용되는 전자문헌을 작성하고 상호 교환하기 위한 지침을 개발하여 더욱 보편적인 언어로 광범위하게 이용할 수 있도록 추진된 프로젝트로서 전문데이터베이스 구축에 있어서 현재 수행되고 있는 인코딩 업무의 복잡성을 줄이면서 전자문헌의 공유를 촉진하고 복잡한 문헌구조를 표현할 수 있는 범용 인코딩 기법을 개발하기 위한 것이다.

안계성/한국데이터베이스진흥센터 정책연구과장

1. 개요

유럽에서는 1960년대부터 인문과학분야의 전자문헌을 각자 독자적인 방법으로 인코딩하여 사용하였다. 그러나 다양한 방법으로 인코딩된 문헌은 네트워크를 통한 정보공유에 부적합하였고 정보의 유통을 위해 다른 형식으로 변환시키는 데에도 많은 시간과 비용이 낭비되었다. 이에 따라 1980년대에 들어서면서 이러한 문제점을 해결하고자 하는 노력이 시도되었다.

TEI(Text Encoding Initiative)는 1987년 뉴욕의 포키프시(Poukheepsie)의 Vassar대학에서 "텍스트 인코딩 지침(Text Encoding Guidelines)"이라는 주제로 열린 회의를 계기로 다양한 인코딩기법을 하나의 코딩 스키마로 통합한 전자문헌의 작성과 교환지침을 제시하기 위한 국제 프로젝트로 시작되었다.

1990년에는 전자문헌의 공유와 기계처리를 위해 복잡한 문헌구조를 기술하는 인코딩 스키마의 개발을 위해 문헌 인코딩과 상호교환 형식을 위한 안내지침서 초안을 출간하였는데, 여기에서는 전자문헌과 문헌유형에 따른 특성을 기술하였다. 1994년 5월 TEI는 기계가독형 문헌(Machine-Readable Text)의 인코딩과 상호교환을 위한 세번째 안인 P3를 발표하였다. TEI는 전문데이터베이스 구축을 위해 국제표준으로 제정된 표준 범용마크업 언어인 SGML을 메타언어로 채택하고 있다.

이와 같이 TEI는 전자문헌을 네트워크를 통해 전송하고 교환하기 위한 표준 상호교환 형식을 제정하고 새로운 유형의 문헌을 인코딩하기 위한 규칙과 원칙을 제시하는 것을 목표로 하여 기존의 인코딩 체제의 문서화와 이를 통해 문헌을 기술하기 위한 메타언어를 제안하고 있다.

2. 사용 환경

TEI는 모든 학문분야의 문헌 인코딩에 대한 규칙을 제정하는 것을 목표로 하기 때문에 TEI 지침은 인문학과 언어학에 기원을 두고 있지만 모든 종류의 문헌을 기술하는데 이용될 수 있는 확장 가능한 구조로 고안되었다.

TEI 구조는 TEI main DTD(Document Type Definition), TEI Lite, Bare Bone TEI로 구별할 수 있는데, TEI main DTD는 전체 TEI 구조를 기술한 DTD이고, TEI Lite는 많이 사용되는 일부 DTD만을 모아서 정의한 DTD이다. Bare Bone TEI는 HTML과 같은 간단한 구조를 가리킨다. 이들 DTD 수준에 따라 TEI 문헌에 포함 혹은 독립되어 존재하는 TEI 헤더(Header)의 내용이 기술된다.

TEI로 작성된 문헌은 TEI 헤더와 텍스트 부분으로 구성되며 〈그림 1〉과 같이 나타낼 수 있다.

TEI 헤더는 상이한 운영 환경에서 사용될 수 있는데 문헌의 일부분으로 포함되어 저자나 출판자가 만드는 경우와 연

```
〈TEI.2〉
    〈teiHeader〉
        〈fileDesc〉 ... 〈/fileDesc〉
        〈encoding〉 ... 〈/encoding〉
        〈profileDesc〉 ... 〈/profileDesc〉
        〈revisionDesc〉 ... 〈/revisionDesc〉
    〈/teiHeader〉
    〈text〉
        〈front〉 ... 〈/front〉
        〈body〉 ... 〈/body〉
        〈back〉 ... 〈/back〉
    〈/text〉
〈TEI.2〉
```

〈그림 1〉 TEI의 구조

구자가 원문을 분석하는 과정에서 헤더를 이용하거나 서지통정의 수단으로 사용되는 경우가 있다.

후자의 경우를 위해 TEI 지침에서는 문헌과 분리해서 저장할 수 있는 헤더인 독립 헤더(Independent Header)에 대한 구조를 규정하고 있다. 독립 헤더는 분산되어 존재하는 TEI로 인코드된 문헌을 참고하여 목록이나 데이터베이스에서 이용할 수 있도록 그 자체가 독립된 구조로 되어 있다.

이밖에도 TEI 독립 헤더는 TEI 형태로 인코딩되지 않은 네트워크 자원을 기술하는 데에도 사용될 수 있으며 이것은 독립 헤더가 메타데이터로 정의되는 배경이기도 하다.

3. 포맷

■ 인코딩(encoding)

TEI 가이드라인은 태그 집합으로 정리된 SGML의 엘리먼트와 속성의 개념으로 원문형태를 정의한다. SGML은 SGML에서 정의한 DTD를 수용한 경우, 개별적인 인코딩을 가능하도록 하는데 SGML을 채택하고 있는 TEI는 DTD의 특정 예로 볼 수 있다.

TEI는 여러가지 선택적인 부가사항을 포함하는 핵심 집합(core set)을 구성하는 확장가능한 구조를 제공한다. TEI내에서 이용되는 태그 집합의 선언을 통해 인코딩되는 문헌에 적합한 DTD를 만들 수 있다.

■ TEI 태그 집합

TEI 문헌은 TEI 헤더 부분과 DTD에 따라 코딩된 텍스트로 구성된다. TEI 헤더는 문헌과 문헌의 인코딩에 관한 정보를 기술한 SGML 엘리먼트 집합으로 표현된다. 따라서 전자문헌의 서지정보와 비서지정보를 수록하는 TEI 헤더에서 순수한 서지정보만은 MARC 레코드와 매우 유사하다고 볼 수 있다.

TEI내의 다양한 엘리먼트는 태그 집합에서 정리되는데, TEI 태그집합요소는 〈표 1〉과 같이 구분할 수 있다.

〈표 1〉 TEI 태그집합요소

Tag 집합	내 용
핵심 집합 (core sets)	모든 문헌에서 필요한 엘리먼트
기본 집합 (base sets)	시, 산문, 드라마와 같이 특정하게 분류된 문헌에 적절한 엘리먼트 집합
추가 집합 (additional sets)	특수하거나 상세한 주제 영역의 문헌 처리에 적절한 엘리먼트
보조 집합 (auxiliary sets)	특수한 역할을 가진 엘리먼트

■TEI 독립 헤더의 내용

TEI 독립 헤더의 구조는 파일 기술(File Description), 인코딩 기술(Encoding Description), 프로파일 기술(Profile Description), 개정내역 기술(Revision Description) 등 4개의 부분으로 구성되어 있다.

이 가운데 파일 기술은 필수 요소이고 인코딩 기술, 프로파일 기술, 개정내역 기술은 선택적인 요소이지만 인코딩 기술은 TEI 독립 헤더 부분에서 특히 권고되는 부분이다.

● 파일 기술(File Description)

TEI 헤더 중 파일 기술(fileDesc)은 일반적으로 정보이용자가 인쇄된 문헌의 표제면에서 찾을 수 있는 서명 · 저자사항, 판 사항, 발행사항, 형태사항, 총서사항 그리고 주기사항과 같은 정보를 포함하며 전자문헌의 원문에 대한 서지정보도 포함한다.

따라서 파일 기술은 도서관 목록, 특히 MARC 레코드형식과 유사하다고 볼 수 있으며 TEI를 기반으로 하는 전자문헌의 파일 기술은 문헌의 기본적인 접근 정보를 제공하

므로 TEI 독립 헤더에서 없어서는 안될 가장 중요한 필수 요소이다. 파일 기술에 포함되는 엘리먼트는 〈그림 2〉와 같다.

〈fileDesc〉			
〈titleStmt〉	.....	〈/titleStmt〉	필수 엘리먼트
〈editionStmt〉	.....	〈/editionStmt〉	권고 엘리먼트
〈extent〉	.....	〈/extent〉	선택 엘리먼트
〈publicationStmt〉	.....	〈/publicationStmt〉	필수 엘리먼트
〈seriesStmt〉	.....	〈/seriesStmt〉	선택 엘리먼트
〈notesStmt〉	.....	〈/notesStmt〉	권고 엘리먼트
〈sourcesDesc〉	.....	〈/sourcesDesc〉	필수 엘리먼트
〈/fileDesc〉			

〈그림 2〉 파일 기술의 엘리먼트

● 인코딩 기술(Encoding Description)

인코딩 기술(encodingDesc)은 인코딩한 문헌의 목적, 표본 추출방법, 편집원칙과 규칙, 인코딩된 문헌과 그 문헌의 원본간의 관계, 원본들간의 관계, 인코딩 수준과 분석수준, 그리고 분류코드에 관한 정보를 기입하는 부분으로 인코딩하는 전반적인 방법에 관한 정보를 상세히 기입한다.

인코딩 기술의 구조는 파일 기술의 경우처럼 간단히 표현할 수도 있고 하위엘리먼트를 세분하여 상세히 구조화시켜 기입할 수도 있다. 인코딩 기술의 엘리먼트는 〈그림 3〉과 같다.

〈encodingDesc〉			
〈projectDesc〉	.....	〈/projectDesc〉	선택 엘리먼트
〈samplingDecl〉	....	〈/samplingDecl〉	선택 엘리먼트
〈editorialDecl〉	.....	〈/editorialDecl〉	권고 엘리먼트
〈tagsDecl〉	.....	〈/tagsDecl〉	권고 엘리먼트
〈refsDecl〉	.....	〈/refsDecl〉	선택 엘리먼트
〈classDecl〉	.....	〈/classDecl〉	선택 엘리먼트
〈/encodingDesc〉			

〈그림 3〉 인코딩 기술의 엘리먼트

● 프로파일 기술(Profile Description)

프로파일 기술은 문헌에서 사용한 언어의 용례, 문헌이나 정보의 생산환경, 그리고 정보생산과정에 참여한 사람이나 기관과 그 구성원에 관한 정보를 기술하는 부분으로서 전자문헌의 검색과 기계적 분석을 위해 필요한 비서지정보를 상세히 기술한다. 또한 전통적인 목록작업에서 가장 중요시하였던 주제분류에 대한 규정을 기입하고 있다.

프로파일 기술은 일반적으로 드라마의 대본, 시나리오와 다중 화자, 목소리를 사용하는 문헌이나 언어학을 기반으로 한 회화문에서 많이 사용한다. 프로파일의 목적은 문헌의 식별과 검색 측면보다 문헌내에는 나타나지 않은 정보를 기술하여 서지정보와 비서지정보를 한 프레임에서 검색할 수 있도록 하는 것이다. 프로파일 기술에 포함되는 엘리먼트는 〈그림 4〉와 같다.

〈profileDesc〉			
〈creation〉	.....	〈/creation〉	권고 엘리먼트
〈langUsage〉	.....	〈/langUsage〉	권고 엘리먼트
〈textClass〉	.....	〈/textClass〉	선택 엘리먼트
〈textDesc〉	.....	〈/textDesc〉	선택 엘리먼트
〈particDesc〉	.....	〈/particDesc〉	선택 엘리먼트
〈settingDesc〉	.....	〈/SettingDesc〉	선택 엘리먼트
〈/profileDesc〉			

〈그림 4〉 프로파일 기술의 엘리먼트

● 개정내역 기술(Revision Description)

TEI 헤더의 마지막 부분인 개정내역 기술은 전자문헌이나 정보가 출판되거나 배포된 후 누가 입력했는가, 교정했는가, 누가 인코딩했는가, 그리고 어떤 시점에서 어떠한 변화가 발생했는가와 같은 변경정보를 기입하여 전자문헌이 생산되고 배포되는 동안 발생한 모든 변화에 대한 정보를 상세히 기술한다.

〈revisionDesc〉				선택 엘리먼트
〈change〉				
	〈date〉	.....	〈/date〉	
	〈respStmt〉	.....	〈/respStmt〉	
	〈item〉	.....	〈/item〉	
〈/change〉				
〈/revisionDesc〉				

〈그림 5〉 개정내역 기술의 엘리먼트

이것은 시스템간이나 연구자들간에 정보파일을 교환·전송할 때 발생하는 파일의 변경정보를 기술하는데 매우 중요한 요소이다. 개정내역 기술의 엘리먼트는 〈그림 5〉와 같다.

구조화된 정보가 적절한 엘리먼트에 포함되면 이들은 AACR2와 ISBD 규칙에 따라 기술된다.

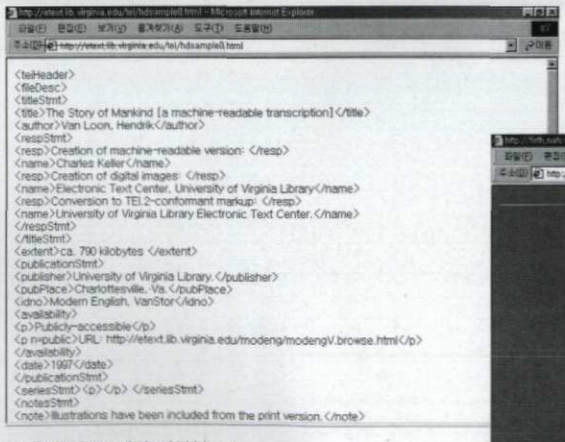
4. 구현 사례

TEI는 주로 인문과학분야의 아카이브를 대상으로 사용되는데 TEI를 이용하여 전자문헌의 데이터베이스 구축과 검색 시스템 구현 사례는 Oxford 대학의 OTA와 Virginia 대학의 전자문헌센터가 대표적이다.

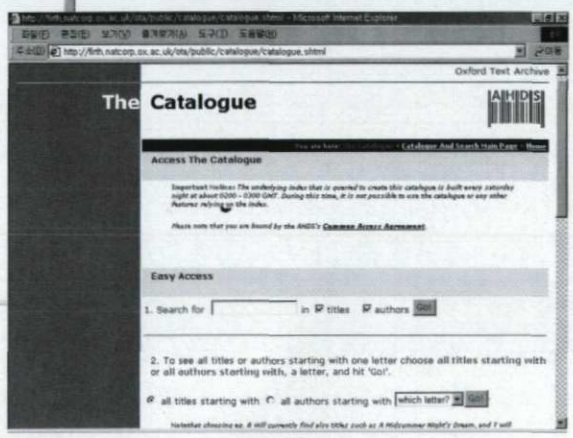
- Virginia 대학의 전자문헌센터(Electronic Text Center) 〈http://etext.lib.virginia.edu/index.html〉

1992년에 설립된 Virginia 대학의 전자문헌센터는 인문학 분야의 전자문헌을 표준형식으로 생산하여 제공하는 역할을 수행하고 있다. 즉, 수천건의 SGML로 인코딩된 전자문헌을 수록한 아카이브를 문헌의 생성과 분석에 적합한 소프트웨어와 하드웨어를 제공하는 도서관 서비스와 접목시켜 제공한다.

현재 12개국 언어로 된 4만 5천건의 텍스트와 관련된 5만개 이상의 이미지가 제공되고 있다. 인쇄 문헌은 Virginia 대학도서관의 목록 서비스를 통해 검색할 수 있다. 데이터 입력을 위한 표준 포맷은 초기에 TEI 지침을 이용하여 구축되었고 현재는 TEI 헤더를 USMARC으로 변환하여 SGML로 생산하고 있다.



〈그림 6〉 TEI 헤더 일부분 예



〈그림 7〉 OTA 목록 검색 화면

전자문헌센터는 지속적인 훈련, 교육, 연구프로젝트에 대한 지원을 통해서 다양한 이용자 그룹을 형성하고 다른 기관의 유사 계획에 있어 하나의 모델로서 제안되고 있다. 〈그림 6〉은 TEI로 작성된 문헌의 TEI 헤더 일부분을 나타낸 것이다.

- Oxford 대학의 OTA(Oxford Text Archive)〈http://firth.natcorp.ox.ac.uk/ota/public/index.shtml〉

OTA는 1976년에 설립된 세계에서 가장 오래되고 잘 알려진 전자문헌센터 가운데 하나로서 Oxford 대학의 전산소(computing service)에서 관리하고 있다. OTA는 현재 25개 이상의 다양한 언어로 된 2,500건의 전자 텍스트와 말뭉치(linguistic corpora)를 소장하고 있으며 이들은 OTA에서 직접 디지털화하여 생산하기보다는 학계 등에 널리 퍼져 있는 고품질의 정보자원으로 엄선하여 수집, 구축한 것이다.

OTA에 수집되는 텍스트의 인코딩은 TEI 지침을 따르도록 하고 있으며 특히 핵심 DTD만을 포함하고 있는 TEI Lite 인코딩 스키마를 사용한다. 〈그림 7〉은 OTA가 현재 소장하고 있는 자료에 대한 목록을 검색할 수 있는 OTA 홈페이지를 나타낸 것이다.

5. 결론

TEI는 학술 연구에 이용되는 전자문헌을 전문데이터베이스로 구축하는데 있어서 복잡한 인코딩 작업을 간소화하고 전자문헌의 상호 교환을 가능토록 하는 범용 인코딩 기법으로 제안된 것이다. TEI는 최근들의 TEI의 장점인 다양한 형태의 응용에 적절한 태그 집합의 개발을 통해 다른 메타데이터와의 호환성을 높임으로써 전자문헌의 전송 및 교환을 위한 국제 표준으로서 자리매김하고 있다.