# TEI Independent Header와 MARC의 비교연구

## A Comparative of TEI Independent Header and MARC

엄혜련 (숙명여자대학교 문헌정보학과 대학원)
김성혁 (숙명여자대학교 문헌정보학과)

Hey Ryen Um, Sung-Hyuk Kim
Dept. of Library and Information Science, Sookmyung Women's University

본 연구는 TEI를 기반으로 한 전자문헌의 서지정보를 수록한 TEI Independent Header를 MARC으로 변환시켜 주기 위하여 전자문헌의 인코딩, 인코딩언어인 SGML, 인코딩 포맷인 TEI를 연구하였다. 나아가 TEI를 기반으로 한 전자문헌의 자동 목록작성의 가능성을 살펴보기 위하여 TEI Independent Header와 MARC을 비교분석하였다.

## 1. 서론

20세기 후반에 등장한 컴퓨터로 인해 기존의 도서관개념이 디지털도서관으로 바뀌어 가고 있다. 이러한 도서관의 환경변화로 인해 인쇄매체의 문헌을 디지털로 변환하게 되었고 도서관은 정보의 공유와 교환을 위한 전문데이터베이스(Full-Text Database)를 구축하기 시작하였다. 네트워크상에 분산되어 있는 이질적인 정보를 이용자가 원하는 장소에서 원하는 정보를 자유롭게 이용할 수 있도록 하기 위하여 전문데이터베이스를 구축해야 한다. 또한 공개, 공유를 목적으로 하는 전문데이터베이스를 구축하기 위해 표준 코딩스키마의 필요성이 부각되었고 이를 위하여 TEI(Text Encoding Initiative)가 시작되었다.

오늘날 정보이용자들은 전문데이터베이스안에 존재하는 수많은 정보를 항해하면서(navigate) 원하는 정보를 검색해야 한다. 그러나 전자문헌의 전문을 대상으로 이용자가 원하는 정보를 검색하는 것은 많은 시간과 노력을 필요로 하기 때문에 기존의 인쇄매체에서 제공해 주던 목록과 같은 검색도구를 전자문헌에도 제공해 줄 필요가 있다. 현재 많은 도서관, 정보저장소, 그리고 이와 관련된 기관에서는 문헌 자체를 수집하지 않아도 네트워크를 이용하여 기계가독형으로 된 문헌정보에 접근할 수 있기 때문에 그 문헌에 대한 서지정보와 그외 문헌정보를 얻을 수 있다. 따라서 이런 기관에서는 원거리 이용자들이 원하는 정보를 검색하고 그 정보에 대한 완전서지정보를 얻을 수 있는 목록, 색인, 데이터베이스를 구축하기 시작하였고 이를 위해 전자문헌의 서지정보를 포함하는 TEI 문헌의 Header에 접근하고자 하였다.

따라서 TEI Independent Header를 기존의 도서관에서 사용해 온 기계가독형 목록인 MARC (Machine Readable Catalog)으로 변환시킴으로 전자정보의 서지정보를 통일시켜 줄 수 있다. 본 논고에서는 전자텍스트 인코딩에 있어서 SGML과 TEI의 역할을 살펴보고, 전자문헌의 서지정보를 수록하고 있는 TEI Independent Header를 MARC으로 변환시키기 위해 TEI Independent Header와 MARC record와의 관계에 대해 살펴보겠다.

본 연구는 전자문헌을 중심으로 한 디지털 도서관에서도 MARC을 이용한 목록이 필요하다는 것을 전제로 하고 있다. 즉, 이는 디지털 도서관시대에도 MARC 표준과 전문의 인코딩 표준이 공존해야 함을 의미한다.

## 2. 텍스트 인코딩

인코딩이란 인쇄된 형태의 정보를 컴퓨터로 처리할 수 있도록 원문을 기계가독형으로(machine-

readable)으로 변환시키는 작업을 말한다. 이는 네트워크 환경에서 정보를 처리하고 공유하기 위해 반드시 필요한 과정중 하나이다. 지금까지의 인코딩은 문헌이 표현하고 있는 다양한 구조와 그에 대한 정보를 무시하는 방법이었다. 따라서 이용자는 'dead document'에서 정보를 탐색하였다고 할 수 있다. 그러나 디지털 도서관시대에는 'live document'인 인쇄매체의 문헌을 그대로 표현할 수 있는 인코딩기법이 개발되어야 한다. 즉 기금까지의 정보검색은 'dead document'이었지만 미래에는 'live document'를 통한 검색이 이루어져야 한다. 인쇄자료의 인코딩 기법은 크게 세가지로 나눌 수 있다. 첫째, 문헌을 이미지로 인식하여 저장매체에 기록하는 이미지 기반 시스템(Image-based system)과 둘째, 문헌을 아스키 문자로 변환하여 전문데이터베이스로 구축하는데 사용하는 아스키 기반 시스템(ASCII-based system), 그리고 마크업언어를 사용하여 원문의 구조를 기술하는 구조화 전문데이터베이스를 구축하는 마크업기반시스템(Markup based system)이 있다. 세번째 방법은 마크업언어로 기술된 문헌이 특정시스템에 한정되지 않는다는 장점을 갖는다 또한 문헌의 구조를 이용한 검색과 효율적인 전문검색을 가능하게 하기 때문에 미국과 유럽의 국가에서는 10년전부터 텍스트 인코딩을 위한 마크업 언어의 개발과 문헌의 구조화를 위한 연구를 추진하여 왔다. 그 결과 텍스트 인코딩을 위한 마크업 언어인 SGML(Information Processign-Text and Office Systems-Standard Generalized Markup Language)을 개발하였으며, 1986년 ISO를 통해 국제표준으로 확정되었다. 또한 국제 표준으로 제정된 표준범용마크업언언인 SGML을 이용하여 문헌의 다양한 구조를 표현하고 인코딩하기 위한 TEI(Text Encoding Initiative)가 결성되어 인코딩 지침을 개발하였다.

## 3. 텍스트 인코딩에 있어서 SGML과 TEI의 역할

### 3.1 SGML

1969년 IBM의 Goldfarb 등은 텍스트의 편집, 포메팅, 그리고 문헌을 공유하기 위한 정보검색 하부시스템을 허용하는 수단으로서 GML(Generalized Markup Language)을 개발하였다. 이후에는 문헌구조, 단축참조, 링크처리, 동시발생 문헌유형과 같은 부가적 개념을 추가하여 1978년에는 GCA GenCode Committee의 후원아래 ANSI에서 SGML의 초안을 작성하였다. 1986년 ISO/IEC JTC 1/SC 18은 ISO 8879 Information Processing Language로 SGML을 표준화하여 문헌구조를 기술하는 메타언어의 표준을 제시하였다. SGML은 문헌의 논리적 및 물리적 구조를 파악하여 문헌의 구조를 이용한 검색과 전문검색, 그리고 원활한 정보 유통을 돕는 것을 목적으로 한다.

SGML 표준은 다음과 같은 사항을 정의하기 위한 수단을 제공한다.

(1) 문헌의 구조
(2) 한 문헌에서 전달되는 문자들
(3) 문헌에서 한 번 이상 사용되는 텍스트
(4) 문헌의 외부에서 생성된 정보를 텍스트로 병합시키는 방법
(5) 텍스트를 구성하기 위해 사용되는 특수한 기술
(6) 텍스트가 처리되는 방법

### 3.2 TEI

유럽지역은 이미 1960년대에 인문과학분야에 컴퓨터를 도입하기 시작하여 텍스트 데이터베이스를 대량으로 집적한 아카이브(archive)나 코퍼스(corpus)를 작성하기 시작하였다. 그러나 이러한 코퍼스나 아카이브는 각각 독자적인 방법으로 문헌을 인코딩한 것으로서 서로 다른 텍스트 인코딩 구조에 의해 텍스트 데이터의 상호교환이나 검색에 있어서 많은 문제점을 불러일으켰다. 이러한 문제점을 해결하고자 텍스트 데이터의 교환과 공유를 목적으로 1987년 "텍스트 인코딩 가이드라인(Text Encoding Guidelines)"이라는 포키푸시회의를 계기로 ACH (Association for Computers and the Hymantities), ACL (Association for Computational Linguistics) , ACH(Association for Compuers and the Humanities), ALLC(Association for Literary and Linguistic Computing)가 공동으로 참여하는 국제프로젝트 TEI(Text Encoding Initiative)가 시작되었다. TEI의 구체적인 목적은 다음과 같다.

(1) 텍스트 데이타베이스 구축에 있어 현재 수행되고 있는 인코딩업무의 복잡성을 줄인다.
(2) 전자문헌의 공유를 촉진한다.
(3) 복잡한 문헌구조를 표현할 수 있는 범용인코딩 기법을 개발한다.
(4) 텍스트데이터베이스의 공통교환포맷의 작성 및 인코딩에서의 문제점 규명하고 이를 해결한다.
(5) 텍스트를 인코딩할 때 어떤 요소가 코딩되어야 하며 또 그것을 어떻게 표현하는가에 관해서 구체적인 조언을 제공하는 가이드라인을 작성한다.

(6) 주된 인코딩법을 조사하여 문헌을 작성하는 동시에 메타언어를 개발한다.

TEI에서는 전문데이터베이스를 구축하기 위해 국제 표준으로 제정된 표준범용마크업언어인 SGML을 메타언어로 채택하여 사용하였다.

## 4. MARC과 TEI Independent Header비교

### 4.1 MARC

현재 대부분의 도서관에서는 서지적 정보를 체계적으로 정리하여 정보를 제공할 수 있도록 구성된 기계가독형 목록인 MARC(Machine Readable Catalog)을 사용하여 왔다. 또한 MARC은 도서관에서 상호 '정보의 공유'를 위해 서지정보를 전달하기 위한 정보교환용 표준형식으로 각각의 내부 형식에서 표준형식으로 변환하여 외부에 서지정보를 보내거나 외부에서 표준형식으로 서지정보를 받아서 내부형식으로 변환하여 사용하기 위한 것이다.

MARC의 모든 레코드는 고정장필드와 가변장필드로 구성되는데 리더와 디렉토리는 고정장필드, 제어필드와 데이터필드는 가변장 필드이다. 리더는 레코드의 처리에 필요한 정보를 갖고 있는 필드이고 디렉토리는 레코드에 출현하는 가변장필드의 정보를 갖고 있는 필드이다. 제어필드는 서지데이터의 처리에 필요한 정보를, 데이터필드는 서지데이터 자체를 의미한다.

### 4.2 TEI Independent Header의 구조

TEI 기반문헌은 전자문헌의 서지정보와 비서지정보를 포함하는 Header부분을 가지고 있다. 이 Header부분은 문헌의 제일 앞부분에 기록되며 문헌의 일부로 취급한다. 그리고 TEI에서는 서지정보를 원하는 이용자에게 문헌을 포함하지 않는 Header를 제공한다. 문헌에 독립적인 Header를 TEI Independenat Header라고 하고 이것은 TEI 텍스트에서 추출한 서지정보와 비서지정보로 텍스트의 일부분으로 간주하지 않고 하나의 독립적인 문헌으로 취급한다. TEI Independent Header에서 file Description은 일반적인 서지레코드를 포함하고 encoding Description, profile Description, revision History은 서지기술의 일부분이나 텍스트분석에 대해 완전 기술정보와 텍스트 인코드방법에 대해 기술한 'codebook'으로 사용된다. 따라서 Independent Header는 도서관, 정보저장소, 개인이 원거리에 있는 기계가독형 텍스트의 서지적정보와 기술적정보, 그리고 완전정보를 얻을 수 있는 주요한 수단으로 제공된다.

TEI Independent Header의 구조는 문헌에 부가된 <teiHeader>와 거의 동일하고 <teiHeader>는 TEI문헌에서 추출하고 필요한 부분의 수정이나 변경을 하면 TEI Independent Header로 사용할 수 있다. TEI Independent Header는 인코드된 텍스트의 완전서지정보, 정보원, 그리고 사용에 있어서의 제한점에 대한 정보를 제공한다. 또한 TEI Independent Header에는 텍스트 그 자체의 인코딩에 대한 정보도 포함한다. 이는 가능한한 완벽한 인코딩 디스크립션을 필요로 한다

#### 4.2.1 File description

TEI Independent Header의 File description은 컴퓨터 파일 자체에 대한 서지적 기술사항이 수록되어 있는 필수요소로 전자파일에 대한 완전서지정보를 포함한다. 그리고 원문이나 전자문헌의 원문에 대한 정보도 포함한다. File description에는 문헌의 知的 내용에 관한 주요책임을 가진 저자, 공저자, 에디터, 편집자와 같은 사람을 header에서 추출하여 그 이름에 대한 적절한 통제(name authority control)를 가한 후 기입한다.

#### 4.2.2 Encoding Description

Encoding Description은 코딩된 전자문헌과 그 문헌을 추출한 원문과의 관계에 대한 정보를 수록한다. <projectDesc>, <samplingDecl>, <editorialDecl>, <refsDecl>엘리먼트는 프로젝트의 과정(process)과 그 목적에 대한 정보를 기록한다. 그리고 텍스트를 어떻게 샘플링하는가, 편집하는 원칙, 그리고 기본적으로 어떻게 참조되었는가에 대한 정보를 제공한다.

#### 4.2.3 Profile Description

Profile description은 문헌의 분류정보나 문맥정보를 기술하고 있다. 특히 텍스트가 어떤 언어로 쓰였는지, 어떤 환경하에서 정보가 생산되었는지, 그리고 생산과정에 참여한 사람이나 기관이 누구인지와 같은 텍스트의 비서지적 정보를 상세히 기술하고 있으므로 TEI Independent Header중 MARC으로 전환시키는데 있어서 가장 문제시되는 필드이다. 이 부분은 검색을 목적으로 사용되거나 텍스트의 machine-supported 분석을 목적으로 사용되고 있는 "codebook"에 로드시킬 수도 있다.

4.2.4 Revision Description

Revision Description은 문헌을 생산하는 동안 파일의 변화와 개정에 관한 정보 및 파일에 대한 모든 정보를 기록해주는 부분이다.

4.3 MARC과 TEI Independent Header의 관계

디지털 도서관은 컴퓨터와 네트워크를 기반으로 하고 있으며, 전세계에 분산되어 있는 정보를 효율적으로 검색하기 위해서는 전자문헌의 목록작업이 필요하다. 따라서 TEI를 기반으로 한 텍스트의 서지적정보와 기술적정보, 그리고 완전서지정보를 원하는 도서관이나 정보저장소, 개인에게 TEI Independent Header를 MARC format으로 전환하여 정보를 제공할 수 있다.

그러나 TEI Indepedndent Header와 MARC사이에는 주요한 차이점이 있다. 바로 각각의 기능이 다르다는 점이다. MARC record는 인쇄매체를 기반으로 하기 때문에 기본적으로 목록카드의 전자버전이다. 목록카드는 복잡한 서지데이터를 포함하는 인쇄매체에 대한 단일 레코드로 구성되어 있다. 반면 TEI Independent Header는 전자매체를 기반으로 발생하였기 때문에 인코딩된 텍스트, 텍스트의 자료원(source), 인코딩 배경(encoding history) 등 TEI 문헌의 완전서지정보를 기록할 뿐만 아니라 Header에 의해 기술된 전자텍스트의 분석을 지원하는 비서지적 정보까지도 포함하고 있다. 그러나 MARC에는 TEI Independent Header에서 제공하는 비서지적 정보에 대한 필드가 없다. 따라서 TEI Independent Header의 profile description, encoding description, revision history와 같이 비서지적 정보는 MARC 레코드안에 여기에 관련된 부분이 없으므로 이러한 내용을 기술하기 위해서는 unstructured note field(5XX)와 같은 필드에 기록해야 한다. Notes field는 보통 기계적 검색이나 분석을 지원하지 않는 반면, 적절히 포맷된 profile, encoding, 그리고 revision description은 각각 검색이 가능하고 기계적 처리가 가능하다. 그리고 Header에 포함된 전자문헌을 직접적으로 지적할 수도 있고 또한 전자문헌은 Header에 있는 관련 엘리먼트에도 지적할 수도 있다.

## 5. 결론 및 제언

디지털 도서관은 정보의 소유개념이 아니라 공유와 접근개념이라 할 수 있다. 따라서 네트워크에 분산되어 있는 이질적인 정보에 대한 메타데이터 작성은 매우 중요하다. TEI를 기반으로 한 전자문헌의 메타데이터의 목록작성은 디지털 도서관의 주요 검색도구로 남아 있을 것이다. 따라서 이용자가 원하는 정보를 효율적으로 검색하기 위해서는 전자문헌의 목록작성이 필요하다. 따라서 TEI 기반문헌의 서지정보가 수록된 TEI Independent Header를 기계가독형 목록인 MARC(Machine Readable Catalog)으로 변환시킴으로 전자 표준목록을 제공하므로 이용자가 원하는 정보를 보다 신속하고 정확하게 검색할 수 있는 수단을 제공해 줄 수 있다.

그러나 MARC에는 전자텍스트의 서지정보를 수록하고 있는 TEI Independent Header에 대응하는 필드는 존재하지만 전자 텍스트의 분석을 지원하는 비서지정보에 대응하는 MARC 필드가 없다. 비서지적 정보는 이에 대한 MARC 레코드 필드가 없으므로 이를 위해 unstructured note field(5XX)와 같은 필드를 만들어 기록해야 한다.

그러므로 전문데이터베이스에 대한 검색을 위한 정보를 제공하기 위해 기존의 MARC format에 다른 형태의 전문(*full text*)을 위한 별도의 format을 연결하여 사용하거나 MARC format을 확장해야 할 필요가 있다. 앞으로 TEI Independent Header를 MARC 포맷으로 자동 변환시킬 수 있는 알고리듬에 대한 연구가 필요하다.

## 참고문헌


김성혁. 1995. '인코딩 포맷 및 방법론.' 1995년 UNION DB 전문가초청세미나 발표자료집 본문 데이터베이스 구축 방법론.

김성혁. 1996. '문헌구조 표현을 위한 표준화에 관한 연구' 한국과학기술원.

Daniel V. Pitti. 1994. 'Standard Generalized Markup Language and the Transformation of Cataloging.'dpitti@library.berkeley.edu.

Jan Corthots. 1995. 'The Use of SGML in the VUBIS-Antwerpen Library Network' jcort@lib.uia.ac.be.

Martin Dillon et. al. 1993. 'Assessing Information on the Internet : Toward Providing Library Services for Computer-Meduated Communication.' OCLC Online Computer Library Center, Inc.

Sperberg-McQueen, C. M. & L. Burnard. 1994. 'Guidelines for Electronic Text Encoding and Interchange. TEI P3.' Chicago : TEI.