효율적 정보자원 공유를 위한 서지 메타데이터 XML DTD 개발

Bibliographic metadata development for the efficient information resource sharing

이혜진, 송인석*
한국과학기술정보연구원 연구원,
한국과학기술정보연구원 선임연구원*

Lee Hye-Jin, Song In-Seok*
Researcher, KISTI,
Senior Researcher, KISTI*

요약

현재 정보 제공 기관들은 분산, 독립적으로 대부분 자료 유형별 서지 메타데이터를 개발, 유형별 검색 서비스 또는 통합 검색 서비스를 제공하고 있다. 그러나 MARC나 MODS등의 관련 표준을 적용한 개별DB스키마의 경우 통합적인 관리나 일관성 있는 유지보수가 어렵고, 특히 자료 유형간의 관계 정의가 되어 있지 않아 전체 자원에 대한 체계적인 자료 수집과 공유체제 구축에는 많은 문제점이 있다.

본 연구에서는 기술적 해결방안으로서 국내외 주요 정보 서비스 기관이 현재 제공하는 학술지, 회의자료 등 여러 자원 유형의 메타데이터 모델을 수집 조사 분석하여 자료 유형별 메타데이터의 체계적 통합 관리를 지원하는 데이터 요소 정의와 모델링 프레임워크 그리고 모듈 기반의 XML DTD를 제안하고자 한다.

Abstract

Most information providers are offering integrated retrieval service based on the bibliographic metadata and schema corresponding to each type of document which are developed in a distributed and independent way. However, it is difficult to maintain the relational consistency of those single heterogeneous databases even though they obey the metadata standard like MARC or MODS. It is the main reason that those standards are restricted to present the general property of document regardless of its type and not to applied to define the relationship of document types. Therefore, It is necessary to define a comprehensive meta model to associate the related databases in a systematic way so that the semantically common part of them can be easily shared and reused without any additional effort like conversion or mapping.

In this paper, we first outline the document types for designing meta model by the empirical analysis of various data schema of main information providers. We propose then data element definition, metadata model and modularized XML DTD which support the efficient and consistent management of multiple ducument types.

Ⅰ. 서 론

인터넷의 증가에 힘입어 정보의 매체가 디지털로 변화함에 따라 초기 디지털 기반 정보서비스 기관은 매체의 변화에 초점을 맞추어 정보를 제공하는 것이 주목적이었다. 분산된 정보환경 속에서 이렇듯 각 자관에 맞는 정보의 디지털화는 메타데이터 표준화의 문제에 봉착하게 되었으며 이를 해결하기 위해 현재까지 많은 노력이 진행되고 있다.

특히 정보서비스 기관의 대표적인 레코드 포맷인 MARC는 디지털객체에 부적합할 뿐만 아니라, 관련 업체와의 교환이나 호환의 폐쇄성, 포맷의 복잡성, 멀티 레벨 레코드의 비생성 등이 문제점으로 지적되어 왔다. 그리고 표준 서지정보 메타데이터로서 기대를 모았던 Dublin Core는 표현요소의 단순성으로 표준의 위상을 확보하는데 실패하였다. 이에 상호운용성과 정밀성을 모두 만족시키고 Dublin Core의 단순성과 MARC의 복잡성을 절충하기 위해 XML을 이용한 MODS(Metadata Object Description Schema)가 개발되었다. MODS는 미국회도서관에서 서지적 요소들을 XML Schema로 표현해서 개발한 메타데이터이다. 이는 디지털 객체의 서지정보 표준 메타데이터로서 다른 유형의 메타데이터의 보완역할과 다양한 속성을 제공하며 외부의 데이터요소 연계가 가능하도록 설계되었다. 그러나 MODS는 다양한 문헌유형에 대한 표현은 가능하지만 유형간의 관계와 특정 문헌이 가지고 있는 특징적 요소 표현에 있어서는 한계점을 가지고 있다. 또한 느슨한(loose) 엘리먼트를 제공하고 있지 않아서 실 시스템에 적용하는 데에는 무리가 따른다.

따라서 본 연구에서는 MODS가 제공하지 못하는 문헌 유형간의 관계와 특정 문헌 요소를 표현할 수 있도록 모듈 기반으로 XML DTD를 설계하는데에 초점을 맞추었다 이를 위하여 문헌유형을 학술지, 회의자료, 단행본, 학위논문, 연구보고서로 정의하고 각 데이터 모델을 수집 분석하여 체계적 통합 관리를 지원하는 데이터 요소 정의와 모델링 프레임워크 그리고 모듈 기반의 XML DTD를 개발하였으며 이는 전체 자원에 대한 체계적인 자료 수집과 공유체제 구축에 그 목적이 있다.

II. 서지 메타데이터 표준

메타데이터는 흔히 데이터에 관한 데이터 혹은 전자자원을 기술하는데 사용되는 데이터로 정의하고 있으며, 최근에는 주로 네트워크 자원의 레코드로 한정된 의미로 사용되기도 한다. 즉, 메타데이터는 데이터에 관한 구조화된 데이터로서 자원에 대한 다양한 접근점을 제공하는 역할을 한다.

서지메타데이터로서는 대표적으로 MARC, Dublin Core, MODS가 있다.

1. MARC

MARC는 39년의 역사를 가지고 있으며, 급변하는 컴퓨터와 통신산업의 변화와 견주어 볼 때, 지속적인 존립은 그 자체로서 상당히 주목할 만하다. 1965년 LC에 의해 MARC I이 완성되었고, 1968년 MARC II가 제정되면서 목록 레코드 구축에 MARC 형식을 최초로 적용하게 되었다. 이러한 MARC는 LC에 편중되어 있었기 때문에 다양한 형태의 정보자료나 통제방식을 지원하도록 확장, 개발되어 왔다. 1973년에는 ISO2709로 채택되어 서지 레코드를 위한 국가 및 국제 표준으로 제안되었다. MARC는 레코드 구조, 내용지시장치, 데이터 내용으로 나뉘어 표현한다.

데이터 자체에 대한 정보는 데이터 필드에 기록되며 데이터 필드는 10개의 영역으로 나뉘어 데이터를 기술하며 유사한 성격의 필드들을 모아준다.

[표 1] 데이터필드 영역

필드	기술 내용
0XX	제어정보
1XX	기본표목
2XX	서명 및 서명 관련 사항
3XX	형태사항
4XX	총서사항
5XX	주기사항
6XX	주제명 부출표목
7XX	부출표목;연관저록
8XX	총서명 부출표목
9XX	자관용 필드

2. Dublin Core

Dublin Core(DC)는 기존 MARC데이터 구조의 경직성과 복잡성으로 인해 네트워크 자원을 기술하는데 많은 어려움이 있어서 이를 극복하기 위한 방안으로 개발되었다. 단순한 구조의 메타데이터 형식을 지원하며 자원의 신속한 검색과 데이터의 호환성 유지를 그 주된 목적으로 한다. 하지만 15개 엘리먼트로 구성되어 서지 기술의 한계를 보이며 또한 디지털 자원 위주의 형식으로 MARC 데이터를 대신하기에는 다소 무리가 따른다.

[표 2] Dublin Core 요소

요소	기술내용
Title	제목
Creator	자원의 책임자
Subject	자원의 주제나 키워드
Description	자원의 내용에 관한 정보
Publisher	출판사
Contributor	지적인 기여자
Date	자원생성날짜
Type	자원의 범주
Format	자원의 데이터표현형식
Identifier	자원 식별자
Sources	해당자원의 출처
Language	자원을 기술한 언어
Relation	다른 자원과의 관계
Coverage	자원의 지리적,시간적 특성
Rights	저작권 관련 내용

3. MODS(Metadata Object Description Schema)

MODS는 MARC의 복잡성과 DC의 단순성을 절충하기 위해 본격적으로 XML을 이용하여 개발된 서지 메타데이터이다. 데이터 요소는 19개의 상위요소와 64개의 하위요소로 구성되어 있다. MODS는 다양한 디지털 자원에 대한 서지정보 기술이 가능하고 디지털 자원의 여러 측면(facet)에 대한 데이터 표현이 가능하지만 현재로서는 적용사례가 적고, 각 자원 유형을 통합할만한 구체적인 지침이 부족하다.

[표 3] MODS의 상위요소

요소	기술내용
Titleinfo	제목관련정보
name	자원의 책임자,기여자
typeOfResource	자원의 범주
genre	자원의 유형
origininfo	출판자 정보와 날짜정보
language	자원 기술 언어
physicalDescription	자원의 데이터표현형식
abstract	초록
tableOfContents	목차
targetAudence	자원이용자
note	주기
subject	주제 및 주제관련키워드
classification	자원의 분류정보
relateditem	관련된 자원
identifier	자원 식별자
location	자원접근정보
accessCondition	저작권관련내용
extension	확장요소
recordinfo	레코드정보

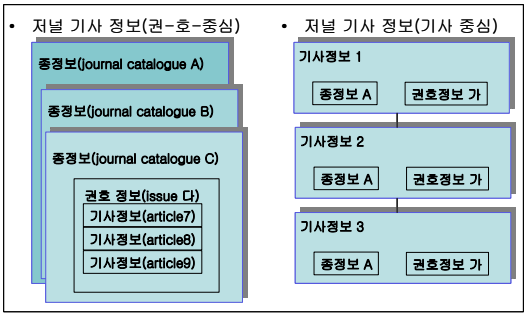
III. 서지 메타데이터 XML DTD

1. 문헌유형별 데이터 모델

본 연구에서는 문헌 유형 별 혹은 DB별 단일접근이 아닌 서로 다른 유형의 정보자원 간의 유기적인 연계를 고려하여 데이터 모델을 분석하였다. 먼저, 문헌 유형은 크게 학술지, 회의자료, 단행본, 학위논문, 연구보고서로 구분지었으며, 단, 단행본의 경우, 문헌의 생성과정이나 기술 될 수 있는 메타데이터 항목이 매우 다양하므로 데이터 모델 분석에서 제외하였다.

1.1 학술지 데이터 모델

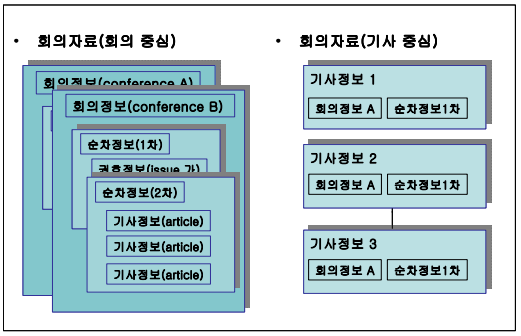
학술지는 종정보 중심과 기사 중심으로 나누어서 데이터모델을 분석하였다. <그림1>처럼 종정보 중심에서는 종정보 내에 다수의 권호정보를 내포하고 있으며 각 권호정보는 다수의 기사정를 내포하는 구조를 가진다. 반면, 기사정보 중심에서 살펴보면 기사정보에 종정보와 권호정보를 함께 내포하고 있는 데이터 모델을 갖는다.



▶▶ 그림 1. 학술지 데이터 모델

1.2 회의자료 데이터 모델

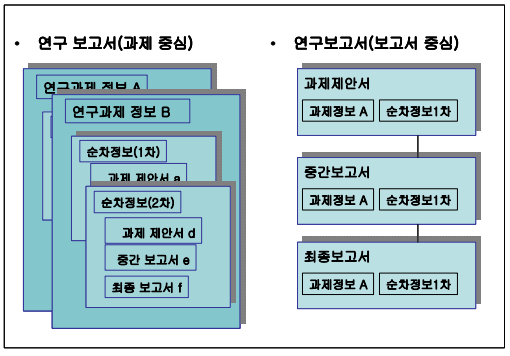
회의자료는 회의 중심과 기사 중심으로 나누어서 데이터모델을 분석하였다. <그림 2>처럼 회의 중심에서는 회의 정보내에 회의순차정보를, 회의순차정보는 다시 기사정보를 내포하고 있다. 기사 중심은 앞서 설명한 학술지 데이터 모델과 유사한 모습이다.



▶▶ 그림 2. 회의자료 데이터 모델

1.3 연구보고서 데이터 모델

연구보고서는 과제중심과 보고서 중심으로 나누어서 데이터모델을 분석하였다. <그림 3>처럼 과제중심에서 살펴보면 과제정보 내에 순차정보를 내포하고, 다시 순차정보에 과제정보제안서, 중간보고서, 최종보고서를 내포한다. 보고서 중심에서는 크게 과제제안서, 중간보고서, 최종보고서로 도메인을 나누고 각 최상위 도메인 내에 과제정보, 순차정보를 내포한다.

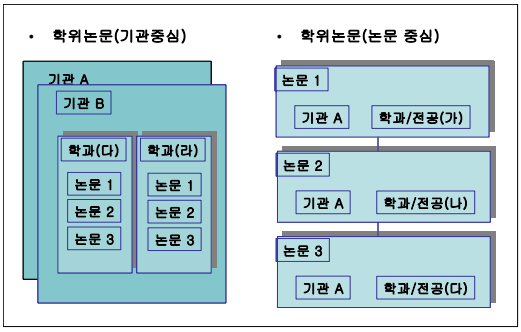


▶▶ 그림 3. 연구보고서 데이터 모델

1.4 학위논문 데이터 모델

학위논문은 기관중심과 논문중심으로 데이터 모델을 분석하였다. <그림 4>처럼 기관중심은 기관정보 내에 학과정보를, 학과정보 내에 논문정보를 내포하고, 논문중심은 논문정보 내에 기관과 학과 및 전공

정보를 내포한다.



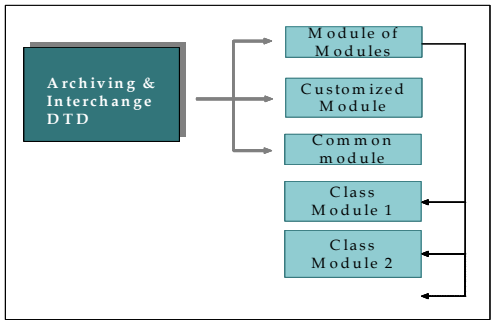
▶▶ 그림 4. 학위논문 데이터 모델

2. 모듈 기반의 XML DTD 문서 디자인

본 연구에서 제안하는 모듈 기반의 XML DTD는 문헌 유형별로 독립된 DTD를 설계하기 보다는 공통항목과 비공통항목(특정 문헌의 특징적 요소)을 기준으로 각각을 구분하여 구성하였다. 즉, 하나의 DTD 내에 여러개의 모듈을 내포하며 이 모듈은 물리적으로 독립된 문서로서 개념적으로 연관성이 있거나 하나의 개념을 구성하는 엘리먼트와 엔터티의 정의를 포함한다. 이는 기존 모듈의 이용 및 맞춤 모듈을 정의하여 자원 유형과 필요에 맞는 효율적이고 상호 운용성이 보장되는 DTD 정의가 가능하다.

모듈구성을 설명하기 앞서 본 연구에서 제안한 XML DTD는 서지정보 메타데이터 기술에 있어 느슨한(loose) 엘리먼트 셋을 개발하는 것을 전제로 한다. 이는 표준을 따르는데 기존 시스템에 무리를 주지 않으며 융통성 있는 메타데이터를 지향한다.

모듈의 구성은 다섯 가지로 구성되어 있다. 첫째, archiving & interchange DTD는 문헌 유형별로 메타데이터 기술에 요구되는 개념단위의 정의를 포함한 모듈의 호출과 기본 데이터 모델을 정의하는 역할을 한다. 둘째, 공통 모듈(common)은 타 모듈에서 참조되는 공통 엘리먼트를 정의한다. 셋째, 모듈 정의 모듈(module)은 DTD 정의에 필요한 모듈을 선언한다. 넷째, 맞춤형 모듈(cutomized)은 특정 목적에 맞게 공통 모듈에서 정의된 엘리먼트나 엔터티를 재정의하거나 새로운 엘리먼트 및 하위 데이터 모델을 정의할 수 있다. 다섯째, 클래스 모듈(meta)은 각 문헌이 갖는 데이터 모델에 따라 엘리먼트의 구조와 카디널리티만 정의해 준다.



▶▶ 그림 5. 모듈간의 관계도

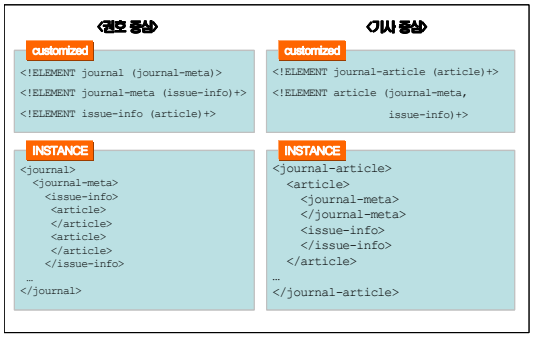
<그림 5>는 앞서 설명한 각 모듈의 관계도를 나타낸 것이다. DTD는 module, customized, common을 내포하며 module은 각 문헌유형에 사용되는 모든 class 모듈을 내포한다. common은 2개 이상의 문헌유형에서 공통적으로 사용되는 엘리먼트와 어트리뷰트, 엔터티를 내포하고 있으며 이렇듯 공통항목을 공유하여 사용하게 되면 유지보수, 통합 검색, DTD 작성의 경제성 등의 장점이 있다. common module이 포함하는 기술 내용은 아래 <표4>와 같다.

[표 4] common 모듈의 기술내용

	기술내용	사용되는 class 모듈
common	서명,키워드,사용언어,format,문헌의 url,목차,소장기관, 저작권정보, 쪽정보, 초록, 문헌수록DB, 발행사항,분류, 창폐간일자, 발행일자, 공헌자정보, 참고사항	모든 클래스
	권호정보, 기사정보, ISSN	학술지, 회의자료
	총서사항, 부속서, 관차사항	학술지, 회의자료, 단행본
	회의정보	회의자료
	ISBN	단행본

본 연구는 DTD 작성시 DTD 작성과 파라미터 및 엘리먼트의 몇가지 유형에 관해 기존방식과 다른 의미로 접근해서 정의하였다. 이는 문헌에 대한 메타정보를 구성하는 개념 및 객체를 기술하기 위한 모델 및 클래스 정의의 경우 실용성과 효율성의 관점에서 고려해야 된다고 보았기 때문이다.

먼저, DTD 정의시, 앞서 데이터 모델의 분석한 결과에 따라 문헌 유형은 어떤 관점에서 분석하느냐에 따라 모델의 구조가 약간씩 다르다. 본 DTD는 다른 메타데이터와는 다르게 이를 수용할 수 있도록 개발되었다. 예를 들어, 학술지의 경우, <그림6>과 같이 customized 모듈을 이용해서 엘리먼트 구조를 권호중심과 기사중심을 모두 수용할 수 있는 DTD를 작성할 수 있다.



▶▶ 그림 6. DTD와 인스턴스 예

엔터티와 엘리먼트 정의는 몇 가지 예로 저자의 경우, 단독, 공동, 기관 등 이질적 구조와 유형의 하위 엘리먼트를 포함하며 경우에 따라 프로파일 정보 중 소속기관 및 연락처의 경우, 반복되거나 공유되어 사용한다. 따라서 엔코딩의 효율성 및 원문 태깅 관점에서 식별자(ID,IDREF) 속성 부여를 해야 한다. 또한 예로 주소의 경우, 지역에 따라 각각 다른 코드체계와 구조를 가지므로 범용적 포괄적 데이터 모델 정의보다는 속성값으로 구분하는 것이 효과적인 방법이다.

3. XML DTD 문서작성 프로세스 및 문서 구성도

이 장에서는 앞서 설명한 각 모듈을 이용하여 문서 작성 과정과 완성된 문서의 구성도를 기술하고자 한다.

문서는 자원 유형별 상위 DTD 구성 요소의 모듈화를 위해 엘리먼트와 어트리뷰트 집합 및 컨텐트 모델은 파라미터 엔터티로 정의하며 최상위 DTD를 제외한 각 모듈문서는 독립된 엔터티 파일로 .ent 확장자를 가진다.

상위 DTD 내에 각 모듈을 호출하되 모듈 정의 모듈, 맞춤형 모듈, 공통모듈, 클래스 모듈 순으로 호출한다.

전체 XML DTD 문서의 구성은 <표4>과 같다.

[표 5] 서지 메타데이터 DTD 구성

서지 메타데이터 DTD				
학술지 DTD	회의자료 DTD	연구보고서 DTD	학위논문 DTD	단행본 DTD
module				
journal customized	proceeding customized	project customized	dissertation customized	monograph customized
common module				
journal meta	proceeding meta	project meta	dissertation meta	monograph meta

VI. 결론 및 제언

XML은 데이터와 문서 교환을 위한 표준 형식으로서 전세계적으로 빠른 호응을 얻고 있다. 특히, 디지털 환경의 급변에 상당히 불리한 상황에 놓여있는 정보 서비스 기관이나 도서관은 이용자의 요구에 맞는 서비스를 위하여 빠른 전환이 필요하다. 하지만 무리한 통합서비스나 디지털 환경에 대한 적응은 이상과 현실의 격차를 넓게만 만들게 될 것이다.

따라서 본 연구는 그러한 단계적인 변화 차원에서 XML을 이용한 서지 메타데이터 표준 DTD를 개발하였으며 이는 기존시스템을 최대한 반영하여 정보

공유의 효율성을 증가시킬 것으로 기대한다. 특히, 각 문헌 유형간의 체계적인 통합관리를 위한 모듈 기반의 DTD는 데이터의 추가, 수정, 삭제 등 관리 측면에서 효율적이고 일관성을 유지하는데 용이할 것이며 새로운 자원 유형의 DTD 설계시, 재사용이 용이할 것이다. 그리고 기존의 MODS나 Dublin Core와 다르게 각 문헌의 고유한 특성을 반영할 수 있어서 표현의 한계성을 극복할 수 있다.

하지만 DTD를 구성하는 모듈의 관계와 구조 파악 및 개별 엘리먼트와 엔터티의 위치파악이 용이하지 않기 때문에 가독의 난해성이 있다. 그리고 현재 다양한 데이터 타입 지원과 DB연계의 용이성으로 각광을 받고 있는 XML schema로의 변환도 수행해야 할 과제일 것이다.

▌참고문헌▐

[1] Gartner, Richarc. 2003. "MODS: Metadata Object description Schema"Libraries and the Academy. Vol. 3. No. 1. pp.137-150.

[2] Heery, Rachel. 1996. "Review of metadata format. Program, Vol. 30. No. 4. pp.345.

[3] Lief Anderson. 2004. "After MARC-what then?" Library Hi Tech. Vol. 22. No. 1. pp.40-51.

[4] Harvad University Library "E-Journal Archival DTD Feasibility Study" 2001

[5] 조윤희 "XML기반 디지털도서관 구현에 관한 연구:XMLMARC 시스템 구축을 중심으로", 제7회 한국정보관리학회 학술대회 논문집. pp.79-82. 2000

[6] 오삼균 "디지털도서관에서의 메타데이터 역할", 한국정보과학회지. Vol. 20. No. 8. pp.45-57. 2002

[7] Journal Archiving and Interchange DTD Tag Library version 1.1 Weg site http://dtd.nlm.nih.gov/tag-library/1.1/index.html [2004.10.4]

[8] MODS Website. http://www.loc.gov/standards/mods[2004.10.4]