디지털 지식정보자원의 상호운용을 위한

메타데이터 활용방안

글 _ 서태설 책임연구원 · 표준화기술지원실 · tsseo@kisti.re.kr

1. 머리말

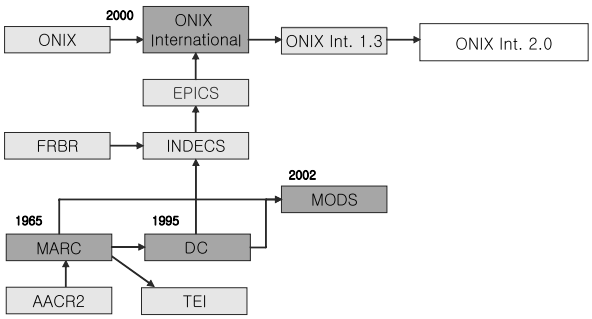
디지털 기술의 발달로 수많은 지식정보자원이 디지털화 되고 있다. 과거에 그러하였던 것처럼 디지털화 된 지식정보도 개발자마다 여러 형태로 가공, 저장, 표현되고 있어 향후 정보의 상호 교환 · 연계 활용에 어려움이 예상된다. 이처럼 정보의 교환이나 공유를 고려하지 않은 정보시스템의 구축과 개별적으로 운영중인 분산된 시스템들은 정보시스템간의 호환성 부재로 지식정보자원의 낭비를 초래하게 될 것이다.

이러한 문제를 해결하기 위한 많은 노력 가운데 가장 주목을 받고 있는 것이 메타데이터(metadata)이다. 이뿐 아니라 최근 W3C(World Wide Web Consortium)을 중심으로 진행되고 있는 시멘틱 웹(Semantic Web) 환경으로 나아가기 위해서도 메타데이터는 온톨로지(ontology)와 함께 빈번히 거론되고 있다. 이 밖에 최근에 새로운 정보유통 패러다임으로 한창 논란이 일고 있는 오픈액세스(Open Access)와 디지털 아카이브(digital archives)의 경우에서도 메타데이터는 화두가 되고 있다.

본 고에서는 이러한 디지털 지식정보자원의 상호운용을 위해서 사용될 수 있는 주요 메타데이터에 대해서 소개하고 MDR을 활용한 메타데이터의 관리와 표준화를 위한 방법에 대해 논하였다.

2. 지식정보자원을 위한 메타데이터

2.1 메타데이터의 정의와 기능



〈그림 1〉 메타데이터의 발전과정

메타데이터는 일반적으로 '데이터에 대한 데이터(data about data)'라고 하며, 대상이 되는 정보자원의 속성과 특성 및 다른 자원과의 관계를 기술하여, 이용자의 관점에서는 검색을 돕고 관련기관에서는 정보원 제어와 관리를 돕는 역할을 한다. 다시 말해서 메타데이터는 실제로 저장하고자 하는 데이터(예를 들면 비디오, 오디오, 텍스트 등)는 아니지만, 이 데이터와 직접적으로 혹은 간접적으로 연관된 정보를 제공하는 데이터를 나타내는 말이다.

메타데이터의 주요 기능은 정보를 식별하고 기술(description)하는 것이다. 그 외에도 정보 객체의 거동 방법, 기능, 사용 그리고 다른 정보 객체와의 관계 및 관리 방법 등에 대한 문서화에도 사용된다. 따라서 메타데이터를 사용하면 사용자가 원하는 데이터가 맞는가를 확인할 수 있고, 쉽고 빠르게 원하는 데이터를 찾아낼 수 있다. 즉, 데이터를 소유하고 있는 측면에서는 관리의 용이성을, 데이터를 사용하고 있는 측면에서는 검색의 용이성을 보장받을 수 있기 때문에 메타데이터의 필요성이 더욱 높아지고 있다.

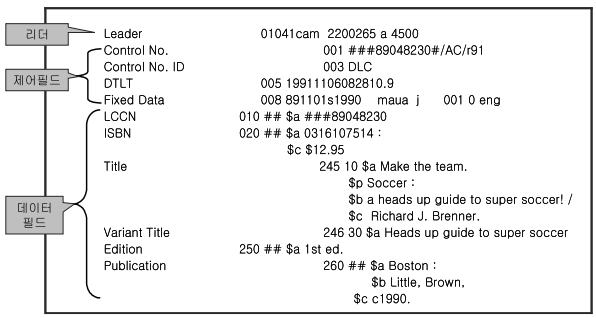
현재 통일된 하나의 메타데이터 표준은 존재하지 않는다. 이는 다양한 환경과 요구에 맞는 적절한 메타데이터 포맷이 요구되며, 이에 따라 각 분야에서 사용되는 메타데이터 구조도 특정 분야의 요구에 맞게 매우 상세하고 전문적으로 개발되고 있기 때문이다. 오늘날 수십 여종의 메타데이터가 사용되고 있으며, 대표적인 것으로 도서관에서 오래 전부터 목록용으로 사용되었던 MARC(Machine readable Cataloging)가 있으며, 전자자원을 기술하기 위한 더블린 코어(Dublin Core) 메타데이터, 온라인 서점과 관련된 ONIX(ONline Information eXchange), 전자도서관을 위한 MODS(Metadata Object Description Schema) 등이 있다. 이밖에도 다양한 메타데이터들이 각각의 목적과 용도에 따라 개발되었고, 또한 개발 중에 있다.

2.2 주요 메타데이터 사례

1) MARC와 더블린코어

1965년에 LC MARC로 출발하여 1973년에 국제표준(ISO 2709)으로 채택된 MARC는 원래 도서관의 목록을 위한 형식을 규정하고 있는 표준으로 세계 각국의 도서관에서 널리 사용하고 있다.

최근 정보가 디지털화 됨에 따라 인터넷 정보자원을 기술하기 위한 메타데이터로서 MARC를 확대 적용하려는 움직임이 있다. 이것은 미국의 의회도서관이 주도하고 있으며 MARC21이라는 이름으로 진행되고 있다.



〈그림 2〉 MARC의 레코드 구조

하지만 MARC는 서지정보를 컴퓨터가 이해할 수 있도록 입력하는 형식으로 오랫동안 사용되어 왔기 때문에 그 구조의 경직성으로 인하여 네트워크 자원을 표현하기에는 많은 비용과 시간이 소요되는 문제가 있다. 이를 대체할 수 있는 단순구조의 형식이 필요하게 되어 새롭게 제안된 것이 더블린 코어(DC)이다.

네트워크 상의 자원에 대한 메타데이터 스키마(schema)들은 각기 다른 요소들로 구성되어 있다. 그러므로 데이터의 호환성을 유지하면서 모든 형태의 네트워크 자원을 기술할 수 있는 일련의 데이터 요소를 규정할 수 있어야 한다.

이러한 목적으로, 1995년OCLC(Online Computer Library Center)와 NCSA(National Centre for Supercomputer Application)가 미국 오하이오주 더블린에서 개최된 워크숍에서 합의한 메타데이터를 더블린코어라고 한다.

요 소	기 술 내 용
Title	제작자(creator)나 발행자(publisher)가 자원에 부여한 제목
Creator	자원의 내용에 책임을 진 개인이나 단체 (예 : 저자(문헌), 화가/사진가/삽화가 등)
Subject	자원의 주제나 그 내용을 기술하는 키워드 혹은 구절
Description	문서의 이미지의 요약 정보를 포함한 자원의 내용에 관한 정보 (예 : 문헌의 초록, 시각자료의 내용기술 등)
Publisher	자원을 현재의 형태로 이용 가능하게 만든 실체 (예 : 출판사, 대학, 기업체)
Contributor	저자이외에 자원의 지적인 기여를 한 인물이나 기관 (예 : 편자, 번역자, 삽화가 등)
Date	자원이 현재 형태로 가능하게 된 날짜
Type	자원의 범주나 장르 (예 : 홈페이지, 소설, 시, 토의문서, 기술보고서 등)
Format	자원의 데이터 표현 형식 (예 : text/html, ASCII 등)
Identifier	자원을 고유하게 식별해 낼 수 있는 문자열 혹은 숫자 (예 : URL, URN, ISBN 등)
Sources	해당 자원의 출처가 된 정보자원 (예 : sources(scheme=ISBN)=0-201-63337-X)
Language	자원의 내용을 기술한 언어 (예 : 영어, 한국어, 독일어 등)
Relation	다른 자원과의 관계. 공식적인 관계를 가지면서 독립적으로 존재하는 자원들과의 관계 표현 (예 : 문서 내의 그림, 책 안의 장(chapter) 등)
Coverage	자원의 지리적, 시간적 특성을 나타내는 요소 (예 : coverage(type=temporal, scheme=yyyymmdd))
Rights	저작권의 사용 권한에 관한 내용. 판권사항이나 판권관리사항, 혹은 이러한 정보를 제공하는 서버와의 연결

〈표 1〉 더블린코어 기본요소

더블린 코어에 대한 확산과 표준화는 DCMI(Dublin Core Metadata Initiative)에서 담당하고 있다. 더블린 코어는 매체의 형식이나 특정 분야, 또는 문화적 배경 등에 상관없이, 모든 유형의 자원에 대하여 기본적인 기술정보를 제공할 수 있는 핵심 메타데이터 어휘를 제시하고자 특별히 설계된 것이다. 즉, 자원탐색을 위해서 사용되는 의미론적 모델만큼은 자원을 담고 있는 매체에 대해서 독립적이어야 한다는 것을 기본 이념으로 삼고 있다.

더블린 코어의 목적은, 데이터의 형식과 구조를 단순화함으로써 원문의 저자나 발행자가 메타데이터를 직접 작성하고, 네트워크 출판을 위한 저작도구의 개발자가 이 정보에 대한 템플릿을 직접 소프트웨어에 포함할 수 있도록 하는 것이다. 만일 특정 주제분야에서 더블린 코어의 15개 기본요소를 상세한 수준으로 확장하려면 한정어(qualifier)를 사용하면 된다. 더블린 코어의 한정어를 캔버라 한정어(Canberra Qualifier)라고 부르는데, 이것은 메일링 리스트를 통한 의견을 문서화 한 것을 기초로 DC Usage Committee의 책임 하에 개발되어 2000년 7월에 발표되었다. 이 한정어는 요소에 대한 한정과 인코딩 스킴으로 이루어져 있다. 요소 한정은 요소의 의미에 대한 확장 없이 요소의 의미를 좀 더 특정적으로 만든다. 예를 들어 'Date' 라는 요소를 한정하여 'Date.Modified' 라는 한정된 요소를 만들 경우, 두 요소 모두 '날짜' 라는 동일한 의미를 가지면서 후자는 '수정날짜' 라는 한정된 의미를 가지게 된다.

인코딩 스킴은 공식적인 외부 표기체계 및 통제 어휘집을 통해 더블린 코어 요소의 값을 용이하게 해석하도록 돕는 참조가 된다. 더블린 코어의 공식 한정어는 최소한의 한정어일 뿐이므로 응용분야에 따른 부가적인 한정어를 정의하여 사용할 수 있도록 하고 있다.

2) ONIX

ONIX는 도서산업의 상품정보를 전자적인 형태로 표현하고 전달하기 위한 메타데이터로 전자상거래 환경에서 도서 유통을 지원할 수 있는 카탈로그 데이터에 대한 요구와 중요성이 증대함에 따라 온/오프라인 통합 유통 표준으로 제안되었다. 1999년 7월 미국출판협회(AAP: Association of American Publishers)가 주관하고 60여개의 출판사 및 도서판매상 들이 참석한 회의에서 ONIX에 대한 개발 논의가 최초로 이루어 졌으며, 미국출판협회가 주최한 회의에서 제안되어 AAP 및 출판사 아마존 등의 전자서점 등이 개발에 참여하였다. 즉, ONIX는 출판업계의 도서정보를 전자적인 형식으로 교환하고 표현하기 위한 표준이다.

2000년 1월에 ONIX Version 1.0이 발표되었고, 최근 2001년 8월에 ONIX Release 2.0이 발표되었다. 계속해서 발표될 Release는 핵심 컨텐츠의 안정성은 유지하면서 표준 범위는 확장될 것이라고 한다. ONIX 프로젝트에는 AAP, EDItEUR, BISG, BIC가 주도적으로 참여하고 있다.

ONIX는 INDECS를 기반으로 하고 있으며, EPICS의 하위집합으로서 XML로 개발되었고, 출판사가 모든 고객에게 단일화된 안정된 데이터 집합을 제공할 수 있도록 지원한다. 그리고 ONIX는 도서에만 국한되지 않고 온라인 거래가 가능한 모든 컨텐츠에 대해 응용 적용할 수 있는 기

회를 제공하며, 목록정보, 저자 정보(저자사진, 음성 등), 도서 정보(책표지, 목차, 크기 등), 출판사 정보, 유통 정보(재고, 구매요청, 가격 등) 등을 처리할 수 있는 현대화 된 개념의 정보를 모두 포함한다.

도서뿐만 아니라 도서 산업을 통해서 출판되고 배포되는 다른 미디어도 포함하는 것을 목적으로 산업의 모든 분야에서 요구하는 특별한 정보를 충족시키나 온라인 서적상(booksellers)으로 제한하지는 않는다. 또한 국가와 세계의 권리, 배포, 가격과 유효성을 실제로 반영할 수 있는 구조를 제공한다.

```
<?xml version="1.0"?>
<!DOCTYPE ONIXmessage SYSTEM
"http://www.editeur.org/onix/2.0/short/onix-international.dtd">
<ONIXmessage>
  <header>  ... </header>
  <product> ... </product>
  <product> ... </product>
  ...
</ONIXmessage>
```

〈그림 3〉 ONIX 메타데이터의 표현 구문 구조

3) MODS

디지털 도서관 영역의 가장 최신 메타데이터인 MODS는 LC의 MARC21 유지보수 부서가 MARC21 이용자 그룹과 함께 만들었다. 2002년에 버전 1.0이 만들어져 6개월간 시험적용한 후 2003년 2월에 버전 2.0이 되었다. 그 후 2003년 9월 현재 버전 3.0까지 나왔다. MODS가 나오게 된 기본적인 요구사항은 XML 환경을 수용할 수 있어야 한다는 것과 MARC21과 최대한 양립할 수 있어야 한다는 것, 그리고 디지털 자원을 수용할 수 있어야 하고 MARC보다는 단순해야 한다는 것이었다.

그 결과로서 얻어진 MODS는 다음과 같은 특징을 갖는다.

* 사용자 편의를 고려한 태그 사용: MARC에서는 숫자 태그여서 의미를 알 수 없었으나, MODS는 단어로 되어 있어서 사용자가 그 의미를 쉽게 알 수 있다.
* 데이터 요소의 그룹화: MARC에서는 별개로 사용되던 태그들이 MODS에서는 그룹으로 표현할 수 있다.
* 코드의 최소화: 널리 알려진 코드의 사용은 허용하였으나 가급적이면 문자열을 그대로 사용하도록 하였다.
* 전자 자원 데이터: MARC에는 없는 전자 자원을 기술하기 위한 태그가 도입되었다.
* 기타: URL과 같은 링킹 정보, 반복의 허용, 특정 속성의 포함, MARC21과의 변환 고려 등의 특징을 갖는다.

Elements	DC
titleinfo	Title
name	Creator Contributor
typeOfResource	Type
genre	-
originInfo	Publisher Date
language	Language
physicalDescription	Format
abstract	Description
tableOfContents	Description
targetAudence	Audience

Elements	DC
note	Description
subject	Subject
classification	Subject
relatedItem	Relation
identifier	Identifier
location	-
accessCondition	Rights
extension	-
recordInfo	-

〈그림 4〉 MODS의 상위요소와 DC 매핑

2.3 메타데이터 상호운용 표준화

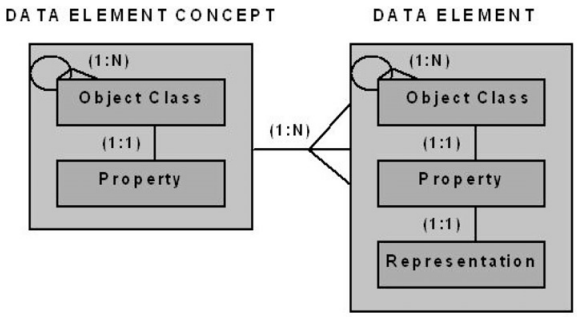
이러한 메타데이터가 폭넓은 이용자 집단에 의해 이용되기 위해서는 많은 관련된 메타데이터들과 조화(accordance)를 이루어야만 한다. 그러나 인터넷의 발전으로 인한 분산된 환경에서 메타데이터를 단일형식으로 통합하는 것은 한계가 있다. 따라서 이용자의 수준과 응용분야마다 요구되는 다양한 데이터 요소를 충족시켜 줄 수 있는 메타데이터의 다양성을 인정하고 이를 수용할 수 있는 메타데이터간의 상호운용성에 관한 연구가 이루어 져야만 한다. 이러한 연구의 대표적인 것으로 워릭(warwick) 구조와 메타데이터 레지스트리(matadata registry)를 들 수 있다.

워릭 구조로서 대표적인 것이 RDF(Resource Description Framework)이다. RDF는 실질적인 적용사례가 아직 적고 XML구문이 아닌 영역에는 적용할 수 없다. 반면 ISO/IEC 11179 표준에 기반한 메타데이터 레지스트리는 구문에 종속되지 않으며, 확장이 가능한 체제이기 때문에 다양한 메타데이터의 상호운용을 하기 위한 적합한 방법이라고 할 수 있다.

메타데이터 레지스트리를 위한 표준으로 ISO/IEC 11179가 있다. 이 표준은 데이터 공유를 위한 기본 틀로서, 데이터 요소(data element)가 내포하는 의미를 충분히 나타내어 줄 수 있는 구조를 정의하고 메타데이터 레지스트리 구축 및 운영 원칙을 제공한다. ISO/IEC 11179는 전체적으로 6개 부분(Parts)으로 구성되어 있다. 제1부(Part 1)는 MDR의 프레임워크를 제시하며, 제2부(Part 2)는 데이터 요소의 분류를 다룬다.

제3부(Part 3)는 MDR의 메타모델(metamodel)과 메타데이터 기본 속성을 제시한다. 제4부(Part 4)와 5부(Part 5)에서는 데이터 요소의 명명, 식별, 정의에 대한 치침을 제공한다. 마지막으로 제6부(Part 6)에서는 데이터 요소 식별자 부여방법과 함께 MDR를 이용하여 데이터 요소를 제안하고 표준상태로 만들어가는 절차에 대해서 규정하고 있다.

ISO/IEC 11179의 1부에서는 메타데이터 레지스트리에 의해 관리되는 데이터의 기본단위에 대한 설명을 다룬다. 그 기본 단위는 데이터 요소이며, 데이터 요소는 3가지 구성요소로 이루어진다(〈그림 5〉 참조).

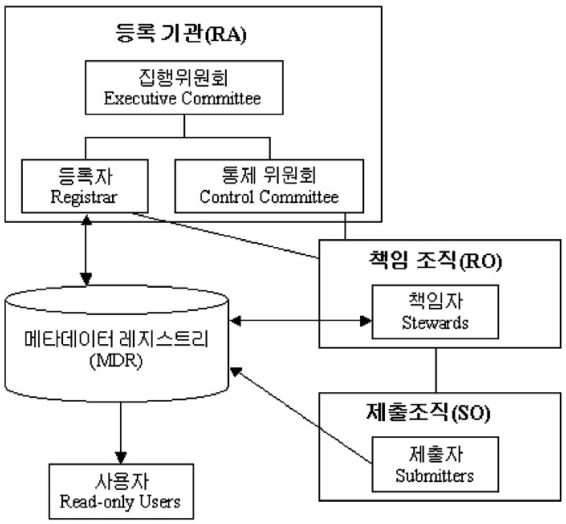


〈그림 5〉 데이터 요소의 구성요소

- 객체 클래스(Object Class) : 실세계의 생각, 추상 또는 사물들의 집합으로 명백한 범위와 의미, 그리고 속성, 행위들이 같은 법칙에 의하여 정의된다.
- 특성(Property) : 한 객체 클래스 내의 구성요소가 가지는 일반적인 특성
- 표현(Representation) : 데이터를 어떻게 표현하는가에 관한 문제로, 값영역(Value Domain), 데이터 타입(Data Type) 등에 해당한다.

ISO/IEC 11179의 3부에서는 공유 데이터의 관리를 위한 메타 모델을 제시한다. 메타 모델은 의미적인 내용과 분산된 환경하의 사용자들이나 정보처리 시스템 간에 공유되는 데이터 요소의 구문을 위한 표준과 안내를 제공하고 있다. 이 표준은 메타데이터 레지스트리의 구조를 개념적 메타 모델의 형태로 명시고 있다. 메타데이터 레지스트리의 구조는 관리 및 식별(Administration and Identification), 명명과 정의(Naming and Definition), 분류(Classification), 관리항목(Administered Item) 등 4개의 영역으로 구성된다.

관리 및 식별 영역에서는 데이터 요소관리를 위한 메타데이터 항목으로 관리항목의 등록자 정보, 식별정보, 제안자 정보, 담당자 정보, 참조정보, 등록/생성 일자, 유효일자, 관련 comment, note 정보 등을 다룬다. 명명 정의 영역에서는 관리항목이 활용되는 주제 분야, 사용언어의 명시, 동의어 등 다양한 표현방법 지원, 언어별 다양한 의미의 사용 지원 등을 다룬다. 분류 부분은 분류 스키마와 분류 스키마 내에 존재하는 저장소 구성요소를 관리하는 데에 사용되는데, 분류체계의 유형 정의, 분류체계를 구성하는 항목, 항목의 관계정의를 다룬다. 관리항목으로 대표적인 것은 데이터 요소, 데이터 요소 개념, 개념 영역, 값영역이 대표적이다. 데이터 요소 개념관리는 데이터 요소의 의미에 해당하는 개념영역에 대한 관리이고, 데이터 요소에 대한 관리는 데이터 요소 자체에 대한 관리이다. 값영역과 개념영역의 관리는 실질적인 표현에 해당하는 부분과 관련된다.



〈그림 6〉 MDR를 이용하는 등록활동 주체

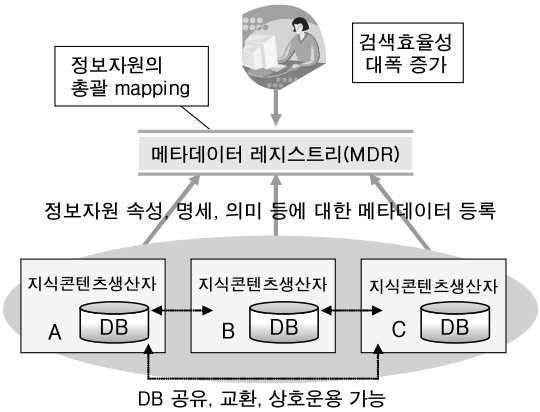
ISO/IEC 11179 제 6부인 '데이터 요소의 등록'에서는 등록자(registrar)를 통하여 데이터 요소를 등록하고, 검증과 인증을 통하여 데이터 표준화를 지원한다. 데이터 요소를 표준화 하는 일은 메타데이터 레지스트리를 중심으로 여러 이해 당사자가 관련이 되어서 이루어진다. 따라서 각 이해 당사자들의 역할과 책임을 분명히 하는 것이 필요하다. 메타데이터 등록 프로세스와 관련된 등록활동 주체(RAB: registration acting body)의 구성은 〈그림 6〉과 같다.

3. MDR을 활용한 메타데이터의 관리

국가의 각종 지식정보들이 여러 곳에서 개발되고 있지만, 호환성 부재 상태로 구축됨으로 인해 정보 이용자들이 정보를 찾는 것이 더욱 힘들어지고, 그에 따라 정보의 신빙성도 확증하기 어렵게 되고 있다. 일례로 연구보고서정보의 경우 정보 공유가 되지 않아 중복 투자가 발생하거나 체계적인 연구관리가 이루어지지 못함으로 국가적인 예산 손실 발생하고 있다. 이처럼 과거의 정보 구축사업에서는 표준화보다는 데이터베이스 구축이 우선적으로 추진되어 타 정보자원(resource)과의 상호 접근이 불가능하거나 제한적인 실정이다. 1990년대 초의 공공DB사업의 경우 정보 표준화를 고려하지 않았으며, 2000년대 초의 지식관리사업도 표준화와 정보개발이 별개로 진행됨에 따라 실효성을 거두지 못하였다.

이러한 문제를 해결하기 위해서는 모든 지식정보들을 표준화된 방식으로 정리 · 관리함으로써 기술정보간의 접근성 · 호환성 · 활용도를 증진할 수 있을 것이다. 이를 운영함에 있어서 단순한 표준화 강조나 표준제정의 수준을 넘어서 각 분야별로 표준을 실제로 활용하도록 하는 장치의 마련이 중요하다. 이를 위해서 국가 단위의 메타데이터센터(MDC: Metadata Center)를 설립하여 효과적인 표준화와 정보유통환경을 마련해 나갈 필요가 있다. MDC에서는 각종 데이터 유형에 따른 메타데이터 표준안을 개발하고, 메타데이터 등록 · 유지 · 관리를 위한 운영지침 마련하며, 이를 교육하고 확산해 나감으로써 국가 지식 자산관리의 효율을 극대화시키게 될 것이다.

메타데이터센터는 이를 운영하기 위한 법제도의 연구도 진행하여야 하며, 국가표준 및 국제 표준화 활동을 통하여 국제사회에서의 국가의 위상을 강화하는 데도 일익을 담당하게 될 것이다.



〈그림 7〉 MDR에 의한 메타데이터 표준화

4. 맺음말

최근 동향으로 볼 때 지식정보에 관련된 다양한 메타데이터는 계속적으로 증가할 것이기 때문에 우리나라도 이 분야의 표준화가 하루빨리 이루어 져야만 한다.

메타데이터를 활용하여 다양한 종류의 정보를 통합 검색, 관리할 수 있어야 하며, 개별적으로 구축된 시스템 데이터간의 상호호환성이 확보되어야만 한다. 그리고 모든 정보의 관리에 공통되는 요소 및 형식에 대한 표준을 제시할 수 있어야 하며, 시스템간의 상호운용과 저작권 관리를 위한 요소들을 추출하여 표준화된 메타데이터와 시스템을 구축할 수 있어야만 한다.

참고문헌

[1] 서태설, 이윤석, 김이란, "지식정보 표준화 핸드북: 21세기 인터넷 시대의 표준과 기술", 한국과학기술정보연구원, 대전, 2001. 12., pp. 130-202.
[2] Susan S. Lazinger, "Digital Preservation and Metadata", Greenwood Publishing Group Inc., Colorado, 2001, p. 143.
[3] 권형진, 지식정보자원 메타데이터, TTA저널, 제78호, pp.61-68., 2001
[4] 김태수, "더블린코어", <http://dewey.yonsei.ac.kr/metadata/DC.htm>
[5] EDItEUR, "ONIX Product Information Guidelines Release 2.0 <Product> record", < http://www.editeur.org>
[6] Sally H. McCallum, "An introduction to the Metadata Object Description Schema(MODS), Library Hi Tech, Vol. 22, No. 1, pp. 82-88., 2004.
[7] 남영광, 서태설, 황상원, ISO/IEC 11179에 기반한 산업기술정보 메타데이터 표준화, 정보관리연구, Vol. 36, No. 1, pp. 57-75, 2005.
[8] ISO/IEC11179 Information Technology - Metadata Registry, <http://metadata-stds.org/11179/>