http://dx.doi.org/10.5392/JKCA.2011.11.9.515

해외 과학기술 학술논문 메타데이터의 비교 분석

Comparison and Analysis of Science and Technology Journal Metadata

이민호*, 이원구*, 윤화묵*, 신성호*, 류재철**
한국과학기술정보연구원*, 충남대학교 컴퓨터공학부**

Min-Ho Lee(cokeman@kisti.re.kr)*, Won-Goo Lee(wglee@kisti.re.kr)*,
Hwa-Mook Yoon(hmyoon@kisti.re.kr)*, Sung-Ho Shin(maximus74@kisti.re.kr)*,
Jae-Cheol Ryou(jcryou@home.cnu.ac.kr)**

요약

글로벌 연구동향 파악, 이머징 시그널 탐지, 선도연구자 파악과 같은 최근의 정보서비스를 지원하기 위해서는 다양한 정보원으로부터 수집되는 대량의 정보를 통합 관리하는 것이 중요하다. 통합 관리를 위해서는 통합 메타데이터 스키마의 정의, 데이터 변환, 스키마 매칭 등의 노력이 필요한데 그 중에서 가장 먼저 수행되어야 하는 통합 메타데이터 스키마를 정의하기 위해서는 현존하는 다양한 메타데이터의 분석이 필요하다. 본 논문에서는 다양한 과학기술 학술논문 메타데이터를 메타데이터의 의미구조, 내용규칙, 구문 등으로 나누어 분석하고 통합 스키마를 만들거나 데이터 변환을 하기위해 고려하여야 할 점을 간략하게 살펴보았다. 일반적으로 구문형태는 편리성과 다양한 사용 환경을 지원하는 XML을 사용함을 알 수 있었으며, 의미구조에서는 공통적으로 사용하는 요소들과 구조화, 계층화한 이름 부여가 필요함을 알 수 있었다. 또한 요소들 중 다양한 내용규칙을 갖는 것들과 관련 표준을 살펴보았다. 분석된 자료는 메타데이터의 통합관리, 데이터 변환, 상호운영을 위한 스키마 매칭 등의 연구에 기초자료로 사용되기를 기대한다.

■ 중심어 : | 메타데이터 | 스키마 | 학술논문 | 정보서비스 | 정보원 | 통합관리 |

Abstract

It is important to manage large amount of information from various information providers for supporting recent information services such as providing global research trends, detecting emerging signal and listing leading researchers. For integrated management, definition of integrated metadata schema, data transformation and schema matching are needed. It is first necessary to analyze existing various metadata for defining integrated metadata schema.

In this paper, we have analyzed several metadata of scientific journal papers by classifying semantics, content rules and syntax, and looked around considerations to make integrated schema or transform metadata.

We have known that XML is used as a syntax for supporting convenience and various usage condition, and hierarchy element names and common elements in semantics are needed. We also have looked at elements having various content rules and related standards.

We hope that this study will be used as basic research material of metadata integrated management, data transform and schema matching for interoperability.

■ keyword : | Metadata | Schema | Research Paper | Information Service | Information Provider | Integrated Management |

* 본 연구는 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단-차세대정보컴퓨팅기술개발사업의 지원을 받아 수행된 연구임(No. 2011-0020516)."

접수번호 : #110803-002
접수일자 : 2011년 08월 03일
심사완료일 : 2011년 09월 08일
교신저자 : 이원구, e-mail : wglee@kisti.re.kr

I. 서 론

최근 정보서비스의 경향은 기존의 단편적인 정보검색 및 제공 서비스를 벗어나 다양한 정보원으로부터 대량의 정보를 수집 및 분석하여 보다 유용한 정보를 제공하려고 하고 있다. 특히 과학기술 분야에서는 과학기술 문헌으로부터 글로벌 연구동향의 파악, 이머징 시그널 탐지, 선도연구자 파악 등을 하기위한 많은 연구가 수행 중이다[1]. 이러한 분석을 위해 수집되는 과학기술 문헌정보는 각 정보원 자신들의 목적에 맞도록 생산되었기 때문에 구성과 표현이 서로 상이하다. 따라서 유용한 정보의 분석 및 추출을 용이하게 하기 위해서는 상이한 정보를 같은 구조와 표현형식으로 통합하여 관리하여야 한다.

통합 관리 및 분석 추출의 중요한 대상이 되는 문헌정보의 메타데이터는 사람에 따라 정의가 조금씩 다르나 본 논문에서는 '어떤 정보 자원에 관한 구조화된 정보를 의미하는 것'을 따르기로 한다[2]. 메타데이터는 기능적인 용도와 의도에 따라 기술용 메타데이터, 관리용 메타데이터, 구조용 메타데이터로 나뉘는데 관리용 메타데이터는 관리를 위하여 관리 기관 내부에서 사용하는 것이기 때문에 수집이 어려우며, 본 논문의 의도는 학술논문을 기술하는 메타데이터를 통합 관리 혹은 데이터 변환을 수행하기 위한 기초자료로 삼기 위한 것이기 때문에 관리 메타데이터와 구조 메타데이터는 본 논문에서 다루지 않는다.

메타데이터 분석과 관련하여 여러 가지 연구가 있었으나, 장기적 보존을 위한 보존 메타데이터의 요소 분석을 수행하였거나 구조적으로 통합하기 위한 구조 메타데이터의 모델링을 한 것이었다[3][4]. 또한 국내 메타데이터 주제의 연구 동향 대부분은 메타데이터 스키마에 대한 이론적 개념소개가 많아 상호운영 방식이나 품질 평가사례에 대한 연구가 필요하다[5]. 따라서 본 논문에서는 다양한 형식의 과학기술 학술 논문의 기술 메타데이터(이하 메타데이터)를 대상으로 비교 분석하여 통합 메타데이터의 설계나 상호운영을 위한 데이터 변환 시에 고려하여야 할 점에 대하여 살펴본다.

분석 대상은 전국 교육기관, 연구기관, 기업체 등의 기관을 대상으로 전자정보의 공동구매를 수행하고 있는 국가 컨소시엄인 KESLI(Korean Electronic Site License Initiative)의 이용자들이 가장 많이 구독하고 있거나 구독을 원하는 학술지를 출판하는 출판사(이하 정보원) 중 메타데이터가 제공되는 것과 구문이나 내용규칙이 타 정보원과 많이 다른 정보원 몇 개를 합해 총 10개를 선정하였다. 참고로 같은 정보원에서 출판하는 학술지의 메타데이터는 대부분 동일하기 때문에 본 논문에서는 정보원을 기준으로 메타데이터를 분석한다.

표 1. 분석 대상 정보원 (학술지)

정보원 (축약명)	설명
Wiley-Blackwell (Blackwell)[6]	- 1807년에 설립된 과학 기술 분야의 선구적인 출판사 - 주제분야 : 화학, 물리학, 법학, 경영, 경제, 교육, 심리학, 컴퓨터학, 생명공학, 의학, 지구과학, 수학, 통계 등
Emerald[7]	- 비즈니스/경영관리 분야의 저널을 제공하는 출판사 - 주제분야 : 경영, 공학, 정보과학, 교육공학
Institute of Physics(IOP)[8]	- 세계적으로 물리학계를 주도하는 IOP학회의 저널을 출판 - 주제분야 : 천문 및 천문 물리학, 생물과학, 화학, 전산과학, 교육, 공학, 재료학, 수학, 계측, 의과학, 나노기술, 물리학 등
JSTOR[9]	- 주요 학술저널의 Archive를 구축하고 광범위하게 이용할 수 있게 확장하는 회사 - 주제분야: 고고학, 식물학, 환경공학, 경영학, 지구과학, 통계학 등
Karger[10]	- 1890년에 설립된 Biomedical Science 분야가 주력분야인 세계적인 출판사 - 주제분야: Biomedical Science
Nature Publishing Group (Nature)[11]	- 전문 과학 및 의료계 서비스가 주 분야인 출판사 주제분야: 의학, 생명과학 등
ScienceDirect (Sciecne)[12]	- 온라인 저널 원문 데이터베이스로 Elsevier에서 발행하는 2천여 종의 저널을 제공 - 주제분야 : 과학 전 분야
Springer[13]	- 1842년에 설립되었으며 전 분야의 우수한 저널 및 단행본을 출판 - 주제분야 : 전 분야
American Chemical Society (ACS)[14]	- 화학 분야의 세계적인 학술기관인 미국 화학학회(ACS) - 주제분야 : 화학
BioOne[15]	- 미 생물학회 및 SPARC (Scholarly Publishing & Academic Resources Council) 협의회 등 도서관 사서, 학자, 학회 및 비영리 출판사들의 상호협력으로 탄생한 비영리 조직 - 주제분야 : 바이오 전 분야

본 논문에서 조사한 과학기술 학술논문 메타데이터 분석 자료와 고려사항은 통합 메타데이터 스키마의 설

계, 메타데이터 간의 데이터 변환을 위한스키마 매핑작업인 크로스워크, 메타데이터 스키마 자동 매칭, 시맨틱 서비스를 위한 기초 연구 자료로 활용될 수 있을 것이다[16-19].

메타데이터의 요소와 특정 목적을 위해 정의된 사용 규칙의 집합을 스키마 혹은 스킴이라고 하는데 본 논문에서는 스키마라는 용어를 사용하며, 메타데이터 스키마에서 명시할 수 있는 메타데이터의 세 측면인 의미구조와 내용 규칙, 구문을 따라 분석한 내용을 기술하였다[2].

2장에서는 각 메타데이터가 어떤 구문으로 이루어져 있는지 살펴본다. 3장에서는 의미구조 측면에서 알아보고, 4장에서는 내용 규칙의 다양함에 대하여 살펴본다. 끝으로 5장에서 결론을 맺는다.

II. 구 문

스키마의 구문은 메타데이터를 이루는 요소들을 어떻게 기계 가독 형식으로 인코딩해야 하는가를 말한다. 메타데이터를 다루기 위해 처리 시스템 내부적으로 표현하거나 저장하는 형태나 타 시스템과 상호교환하기 위한 형식이다. XML이나 SGML과 같이 구문은 메타데이터의 의미구조와 얽혀 있는 채로 정의되는 경우도 있고, 스키마 자동 매칭이나 메타데이터 변환을 위한 파싱에서 사용되기도 한다. 그러므로 메타데이터가 어떤 형태 혹은 형식으로 표현되어 있는가는 우리의 목적 달성을 위하여 꼭 알아야 할 필요가 있다.

표 2. 메타데이터 정보원별 구문형태

정보원	구문 형태
ACS	xml
BioOne	tagged text
Blackwell	xml
Emerald	sgml
IOP	xml
JSTOR	xml
Karger	xml
Nature	xml
Science	sgml
Springer	xml

[표 2]에 각 정보원별 메타데이터의 구문형태가 표시되어 있다. Tagged text는 정보원에서 자체 정의한, 요소를 구분할 수 있는 태그를 일반적인 텍스트로 기술한 형태를 말한다. Tagged Text 형식인 BioOne의 메타데이터 일부가 [그림 1]에 표시되어 있다. 논문의 제목은 'TI', 저자는 'AU', ISSN은 'IS'로 요소이름을 표시하고 ':'을 구분자로 하여 메타데이터를 표시하고 있는 경우이다. [표 2]에서 볼 수 있듯이 Tagged Text와 SGML을 사용하는 경우도 있지만, 대부분의 정보원들이 XML을 메타데이터 구문(형태)로 많이 사용하고 있는 것을 알 수 있다. XML은 SGML이 가지는 프로그램이 처리하기 어렵다는 심각한 단점을 없애기 위하여 SGML보다 더 엄격한 규칙을 적용한 SGML의 부분 집합이다. 또한 웹상에서 사용하기 더 적합하며 국제적인 환경에도 더 잘 맞고 많은 브라우저와 애플리케이션이 지원하고 있다. 이러한 편리성과 사용 환경에 따라 각 정보원들에서도 XML을 메타데이터 구문으로 많이 채택하고 있다고 볼 수 있다. 여러 정보원의 데이터 통합을 위해서는 각 구문을 해석할 수 있는 여러 가지 파서를 모두 갖추어야 하지만, 조사한 결과를 토대로 보면 XML 파서만 가지고도 대부분의 메타데이터를 해석할 수 있음을 알 수 있다.

```
  #1
  TI: Compatibility Relationships in Distylous Bluets
– Houstonia serpyllifolia and H. longifolia
(Rubiaceae)
  AU: Beliveau, BD; Wyatt, R*
  AF: Highlands Biological Station,
P.O.Box580,Highlands,NC28741,USA
  SO: American
MidlandNaturalist[Am.Midl.Nat.].Vol.141,no.2,pp.21
7–226,Apr1999.
  IS: 0003–0031
  .... 이하 생략
```

그림 1. BioOne Tagged Text의 예

III. 의미구조

의미구조는 메타데이터의 요소 자체의 의미를 말하

는 것으로 일반적으로 이름과 정의로서 의미를 표시한다. 서로 다른 정보원이 작성하는 메타데이터는 같은 의미를 갖는 요소를 다른 이름으로 정의하거나 반대로 같은 이름을 다른 의미로 사용하는 경우가 많이 있다. 또한 요소들 여러 개를 그룹으로 묶거나 요소들의 상하위 관계가 있을 경우 메타데이터 스키마들 간에는 순서나 배열 등의 구조적 차이가 있을 수 있다. 의미구조를 파악하는 일은 스키마 크로스워크나 스키마 자동매칭을 위해서 꼭 필요하다.

과학기술 학술논문을 기술하는 메타데이터들에게서 공통적으로 사용되는 요소들은 [표 3]과 같다. 요소들을 살펴보면 주로 논문검색 후 나오는 간략보기 화면에서 볼 수 있는 요소들은 전부 사용하고 있음을 알 수 있다. 의미상으로 거의 비슷한 요소들이기는 하나 실제 이름과 구조는 조금씩 다르다.

표 3. 10개 정보원 메타데이터의 공통요소와 이름

요소	요소 이름 (Nature)	요소 이름 (JSTOR)
작성 언어	Language	tei.2/text/body/div@lang
초록	Abstract	tei.2/text/body/div/bibl/p@type
저자	AuthorList/Author/FirstName, AuthorList/Author/MiddleName, AuthorList/Author/LastName	tei.2/text/body/div/bibl/author@type
페이지	FirstPage, LastPage,	tei.2/text/body/div/bibl/extent
출판일	Journal/PubData/Day, Journal/PubData/Month, Journal/PubData/Year	tei.4/teiheader/filedesc/publicationstmt/date
논문 제목	ArticleTitle	tei.2/text/body/div/bibl/title@variant
권호	Journal/Volume	tei.2/teiheader/filedesc/publicationstmt/idno@type
ISSN	Journal/Issn	tei.2/teiheader/filedesc/publicationstmt/idno@type
저널 (논문지) 제목	Journal/JournalTitle	tei.2/teiheader/filedesc/titlestmt/title@type

작성언어의 경우를 보면 대부분의 정보원들은 별도의 요소이름으로 기록하고 있으나 JSTOR는 속성으로 기록하는 경우도 있다. 저자나 출판일 처럼 같은 요소를 세분화하여 여러 요소이름으로 나누는 경우도 있다. 특히 초록이 가장 많은 차이를 보였는데 ACS와 Springer와 같은 경우 상당히 세밀하게 요소를 나누어 기록하고 있었다.

또한 속성을 이용해 요소에 추가적인 의미를 부여하는 경우도 있었다. 대표적인 정보원이 JSTOR인데 저자에 type 속성을 이용해 주 저자, 교신저자, 공동저자를 구분하고 있었으며 초록의 경우에도 type 속성을 이용해 저작권 정보까지 기술하도록 하고 있었다.

표 4. 같은 의미이나 구조가 다른 요소들

정보원	내용
ACS	초록의 세분화 (그림요소까지 포함) 저널 축약제목 존재
BioOne	초록 언어 선택가능 저널 축약 제목 존재
BlackWell	출판일 요소에 연도, 월까지만 존재
IOP	저널 축약제목 존재
Karger	저널 축약제목 존재
Science	저널 축약 제목 존재
Springer	초록의 세분화 저널 축약제목 존재

[표 4]는 공통요소 중에서 정보원별로 구조나 의미가 약간 다른 것들을 더 자세히 나열한 것이다. BioOne의 경우 초록을 기술하는 언어도 기재할 수 있도록 되어있다. 또 대부분의 정보원들이 출판일은 연도, 월, 일까지 기술할 수 있는 요소가 별도로 있거나 하나의 요소에 연월일을 모두 기재하는데 반해, BlackWell의 경우는 일을 기재할 수 있는 요소가 없다. 대부분의 정보원들이 학술지명(저널)의 축약제목을 기재할 수 있는 요소를 갖는다는 점도 흥미롭다. 이것은 각 정보원별로 갖는 추가적인 요소이나, 의미상으로 같은 내용을 조금 다르게 표현하는 요소를 갖는 것이기 때문에 의미상 공통요소로 분류하였다. 위와 같이 의미의 일부누락(출판일의 일자 누락)이나 구조상의 차이(초록의 세분화)와 같은 것들은 스키마 통합에서 많은 합의가 이루어져야 하는 부분이며, 스키마 자동 매칭 연구의 가장 큰 주제이기도 하다.

표 5. 정보원별 추가요소

예외 요소	설명	비고
기사 타입	학술지, 프로시딩인지 종류를 표기	
ISBN	프로시딩인 경우 사용하는 ISBN 번호	
분류	기사 주제에 따른 분류 표시	
저자 소속기관	저자의 소속기관	
저자 이메일	저자의 이메일	
소속기관 주소	저자 소속기관의 주소	국가명을 분리하는 경우도 있음
통권 (issue)	창간호부터 현재까지 발행한 권수 혹은 해당 권에서 발행한 호수	
저작권	저작권 소유자, 기간, 설명 등	
발행형태	온라인, 오프라인 발행표시	
원문링크	온라인 발행인 경우 원문으로의 링크	DOI로 기술하는 경우가 일반적이나 그렇지 않은 경우도 있음
발행기관명	발행기관명칭	
발행기관 주소	발행기관의 주소	발행기관의 국가명, 도시명을 분리하는 경우도 있음
부제목	논문의 부제목 또는 학술지명의 부제목	부 제목의 언어를 표시하는 경우도 있음
참고문헌 수	논문에 기술된 참고문헌의 수	
키워드	논문 검색에 사용할 키워드	
페이지 수	논문 페이지의 수	
계절	발행계절 표기	

[표 5]는 공통요소 외에 각각의 정보원들만이 가지는 추가적인 요소들이다. 통합 메타데이터 스키마를 작성할 때는 응용 분야의 필요에 따라 해당 요소들을 더 포함할 것인지를 결정하여야 한다. 모든 정보원들을 아우르기 위해서는 상기 요소들을 모두 포함하는 메타데이터 스키마를 설계하여야 할 것이다. [표 5]에서 나열하는 요소들 중 요소의 내용이 달라 통합이 쉽지 않은 경우도 있다. 대표적인 것이 분류 요소이다. 분류는 도메인별로 다르며, 같은 도메인일지라도 사용하는 분류체계가 달라 어려움을 겪는 경우가 많다. 이러한 경우 분류체계의 통합이 선행되어야 한다. 통권의 경우도 사용처마다 조금씩 다른 의미로 사용하는 경우가 많아 주의하여야 하는 요소이다. 페이지 수 요소는 공통요소인 페이지 요소로부터 추출할 수 있기 때문에 데이터 변환이 비교적 쉬운 요소라 할 수 있다. 부제목의 경우는 논문 제목의 부제목을 기술하는 경우도 있으나 일부 정보원에서는 영어가 아닌 논문의 경우 원어 제목을 부제목에 기술하고 영어 제목을 공통요소인 논문제목에 기술하고 있었다.

IV. 내용규칙

내용 규칙은 메타데이터 요소의 값이 어떻게 선정되고 표현되는지를 명시한다. 내용규칙은 요소의 값을 결정하는 방법, 값을 표현하는 포맷, 취할 수 있는 값의 집합이나 범위를 명시할 수 있다. 이것 또한 메타데이터 작성자별로 상이하다. 예를 들어 2개의 서로 다른 정보원에서 논문의 저자를 author라는 이름으로 하나의 요소를 사용하여 XML로 표시한다고 할 때, 두 스키마에서 저자 요소의 의미구조와 구문은 같다고 할 수 있다. 하지만 내용규칙이 다르다면, A라는 학술지에서는 '성, 이름' 으로 표시하고 B라는 학술지에서는 '이름, 성' 으로, C라는 학술지에서는 '이름. 성.' 으로 표시할 수도 있다. 이것은 데이터 변환이나 상호운영에 있어서 걸림돌이 될 수 있다. 일반적인 해결 방법 중의 하나는 시소러스나 전거파일 또는 용어리스트와 같은 통제어휘를 사용하거나, 공통의 분류코드, 식별자등을 사용한다. 사용된 내용규칙을 정확히 아는 것은 다른 메타데이터로의 변환 과정에서도 필수적이다. 내용 규칙은 정보원이 작성한 요소의 세부 설명을 보거나 실제 데이터를 면밀히 조사하여 파악하여야 한다.

[표 6]은 각 메타데이터에서 서로 다른 내용 규칙을 사용하고 있는 것들을 나열한 것이다. 논문 검색이나 참조를 위해서 논문 제목, 저자, 출판사, 저널 제목, 권호, 페이지 번호는 반드시 필요한 요소들인데, 그 중 페이지 번호, 권호, 저자 요소들은 시스템으로 처리하기 위해서 반드시 공통된 내용 규칙으로 통일되어야만 한다. 그 외 각 제목이나 초록과 같은 요소를 나타내는데 있어서도 사용되는 특수문자 코드 셋이 달라 상호호환이나 통합에 있어서 문제가 되기도 한다.

표 6. 상이한 내용 규칙과 영향을 주는 요소들

내용 규칙	설명	해당되는 요소이름
페이지 표기	페이지 수를 표기하는 규칙	페이지, 페이지 수
사용 언어 표기	논문이 기술된 언어를 표기하는 규칙	사용언어, 초록언어 등
논문 형태	논문이 학술지인지 프로시딩인지를 표기하는 규칙	기사타입, 발행형태
원문 링크	원문 URL 표기 규칙	원문링크
권호 표기	권호 표기 규칙	권호, 통권
저자 형태	저자의 기여 형태 (주저자, 교신저자, 공동저자 등)	저자명
날짜 표기	연,월,일 표기 규칙	출판일, 발행일
분류	분류체계, 명칭, 코드 등의 차이	분류
저자명 표기	저자명 표기 규칙	저자명
계절 형태	계절 표기 규칙	출판일, 발행일, 계절
특수 문자	수학기호, 도형문자, 움라우트 표기 등	초록, 제목 등

계절별로 발행되는 계간지의 경우 출판일 또는 발행일을 표기하지 않고 봄, 여름, 가을, 겨울과 같은 계절명을 기록하는 경우도 있다. 이러한 계절명을 표기하는 규칙도 상이하므로 일관되게 변경할 필요가 있다. 다음은 논문 검색이나 참조를 위하여 반드시 통일되어야 하는 요소의 내용 규칙들을 자세히 알아본다.

■ 저자명 표기 규칙

저자명 표기 규칙은 일반적으로 XML 혹은 자체 정의한 태그를 통해서 이름과 성을 다른 항목으로 구분하여 관리하는 정보원들이 대부분이다. 하지만 하나의 항목에 같이 표기하는 경우도 있는데, 이런 경우 성과 이름의 순서가 다르기도 하고, 이니셜로 축약하는 경우와 이름 전체(Full name)를 풀어서 사용하는 경우 등 세부적인 표기에 있어서는 매우 다양하다. 이름을 이니셜로 표기하는 경우에도 이니셜 사이에 대시(-)를 넣거나 성과 이름 사이에 컴마(,)를 넣는 경우도 있으며, 빈 공백으로 구분하는 경우도 있다.

표 7. 다양한 저자명 표기 규칙

정보원	표기 규칙
BioOne	– "성, 이름"의 형식으로 표기됨. – 저자와 저자는 "세미콜론블랭크(;)"로 구분 예) AU: Amer,MA;Miura, Takeshi*; Miura, Chiem
ACS	– 소스데이터 구조: "성, 이름[, 가계표시어]" 예) 〈AUI〉Baxter, P. J.〈/AUI〉 〈AUI〉Howell, T. A., Jr.〈/AUI〉
Karger	– DTD : Author (initial|forename?|surname)+ 예) 〈author authtype="au"〉 〈initial〉A.〈/initial〉 〈surname〉Lazar〈/surname〉 〈/author〉 〈author authtype="au"〉 〈initial〉S.C.C.M.〈/initial〉 〈forename〉Silvia C.C.M.〈/forename〉 〈surname〉van Coeverden〈/surname〉 〈/author〉
Nature	– DTD : Author ((((FirstName, MiddleName?, LastName, Suffix?)| CollectiveName),Affiliation?) 예) 〈Author〉 〈FirstName〉W〈/FirstName〉 〈MiddleName〉L〈/MiddleName〉 〈LastName〉Lowe〈/LastName〉 〈Suffix〉(Jr)〈/Suffix〉 〈Affiliation〉Columbia...〈/Affiliation〉 〈/Author〉 〈Author〉 〈CollectiveName〉〈/CollectiveName〉 〈Affiliation〉University...〈/Affiliation〉 〈/Author〉

[표 7]은 다양한 이름 규칙을 사용하고 있는 정보원들을 보여준다. BioOne처럼 하나의 요소에 다수 저자가 ';'을 구분자로 표기되는 경우, ACS처럼 가계 접미어를 이름 요소에 붙여 사용하는 경우에는 구분자를 잘 살펴야 데이터 변환 시 분리해, 각각의 요소를 추출할 수 있다. 반대로 분리된 요소를 하나의 통합요소로 기술하고자 할 경우에는 저자 이름 축약 시 사용하는 마침표(.)와 같은 기호와 혼동되지 않도록 잘 선택하여야 한다. 또한 Nature의 예처럼 단체 저자는 새로운 요소를 만들어 표기할 것인지, 표기 규칙에 대한 정의도 필요할 것이다.

■ 권호 표기 규칙

권호 표기 규칙은 일반적으로 정보원마다 가장 많이 차이가 나는 항목이다. 따라서 권호 표기 규칙에 대한 조사와 표준화가 이전부터 많이 진행되어 왔다. 국내의 경우에는 2004년 과학기술 정보표준화위원회에서 권호

표기 규칙을 조사하고 표준화하여 사용을 권고하였다[20]. [표 8]과 [표 9]는 과학기술정보표준화위원회에서 조사한 권호 명칭의 다양한 표기 규칙들이다. 본 논문에서 조사한 정보원들의 메타데이터의 경우에는 요소명의 차이가 있기는 하였으나 비교적 권호 정보가 분리된 요소로 잘 표현되어 있으며, 표기규칙에서도 큰 차이는 없었다. [표 10]은 조사한 정보원들 중 일부의 권호 표기 규칙의 예이다.

표 8. 권호 명칭의 다양한 유형

대표명칭	유형
Volume	권, volume, VOLUME, v, V, Vol, volumes 등
Number	호, number, Number, n, N, no, NO, num 등
Issue	Issue, issues, issue no 등
Part	편, part, PART, P, PT, pts 등
Supplement	부록, supplement, suppl, supple. No. 등

표 9. 권호 넘버링의 유형

넘버링 유형	표기 예
숫자	v.11 no.3
문자	v.117 no.7 pt. A
[숫자][문자]	v.9A no.8
[문자][숫자]	v.4 no.A4
문자-숫자	v.ASSP-22 no.1

표 10. 권호 표기 규칙의 예

정보원	표기 예
Nature	〈Volume〉79〈/Volume〉
BioOne	JV: 141 JI: 2
Springer	〈VolumeInfo VolumeType="Regular"〉 〈VolumeIDStart〉37〈/VolumeIDStart〉 〈VolumeIDEnd〉37〈/VolumeIDEnd〉 〈VolumeIssueCount〉6〈/VolumeIssueCount〉 〈/VolumeInfo〉

통합 스키마를 설계할 때에는 향후 상호운영에 문제가 없도록 권호 정보 표준 표기를 따라 내용 규칙을 잘 정의하여야 할 것이다.

■ 페이지 표기 규칙

페이지 표기는 학술지에 따라 표기 규칙이 매우 다양하다. [표 11]에서 보듯이 같은 정보원이라도 숫자로 쓰기도 하고 로마자로 쓰기도 하는 등 페이지 표기를 다양하고 하고 있었다. 다르게 표기하기는 하나 일정한 규칙을 추출하기가 매우 쉽고 일반적으로 의미가 매우 명확하며 구조가 단순하기 때문에 스키마 매칭이나 데이터 변환이 다른 요소에 비해 쉬운 편이다.

표 11. 페이지 표기 규칙의 예

BioOne	Nature
JP: 1	〈FirstPage〉v〈/FirstPage〉 〈LastPage〉vii〈/LastPage〉
JP: 1-1	〈FirstPage〉S139〈/FirstPage〉 〈LastPage〉S140〈/LastPage〉
JP: 1-396	〈FIRSTPAGE〉BE67〈/FIRSTPAGE〉 〈LASTPAGE〉BE69〈/LASTPAGE〉
JP: 001-002	
JP: 10S-12S	
JP: 999-1012	
JP: II-XI	
JP: p. 10	
JP: p. 106S	

■ 언어, 논문 형태 내용 규칙

값이 한정될 수밖에 없는 언어나 논문 형태는 코드셋을 이용하는 경우가 많기 때문에 데이터변환을 하기가 용이한 편이다. [표 12]는 언어 코드의 예를 보여주고 있다.

조사한 정보원들의 대부분은 ISO639 코드나 USMARC 코드 등 표준 코드를 사용하고 있었다. 하지만 BioOne의 경우처럼 일부는 코드를 사용하지 않고 언어이름을 그대로 사용하는 경우도 있었다. 언어나 논문 형태와 같은 경우에는 어느 정도 값이 한정되어 있기 때문에 코드와 같은 통제어휘를 사용하는 것이 오타와 같은 잘못된 입력이나 이용자마다 다른 어휘를 사용하는 경우에 강인할 수 있어 바람직하다 할 수 있겠다. 하지만 코드 테이블에 없는 언어의 경우 표기할 수 없다는 단점도 있다. 예를 들어, 본 논문의 조사 대상에는 들어있지 않았던 어떤 정보원의 경우 영어, 프랑스어, 독일어를 제외한 나머지 언어들을 구분하지 않고 하나의 코드 값으로 표기하는 경우가 있었다.

표 12. 정보원들이 사용하는 언어 표기 규칙의 예

언어	Nature	BioOne
English	EN	English
French	FR	French
German	DE	German
Russian	RU	Russian
Spanish	ES	Spanish
Italian	IT	Italian

V. 결론

최근의 정보서비스를 지원하기 위해서는 다양한 정보원으로부터 수집되는 대량의 정보를 통합 관리하는 것이 중요하다. 이를 위해서는 서로 다른 메타데이터 스키마로 표현되는 메타데이터의 통합 스키마의 정의, 데이터 변환, 스키마 매칭 등의 노력이 필요하다. 본 논문에서는 과학기술 학술논문에서의 여러 메타데이터를 조사하였다. 메타데이터의 의미구조, 내용규칙, 구문 등으로 나누어 분석하고 통합 스키마를 만들거나 데이터 변환을 할 때 고려하여야 할 점을 간략하게 살펴보았다. 살펴본 바에 따르면, 관리 혹은 서비스의 목적에 따라 달라질 수는 있겠으나, 일반적으로 구문형태는 다수의 정보원이 사용하며 편리성과 다양한 사용 환경을 지원하는 XML을 사용하는 것이 더 포괄적인 것임을 알 수 있었다. 의미구조에서는 [표 3]에서 제시한 공통요소는 반드시 포함하되 이름에서 요소의 의미를 파악할 수 있도록 구조화, 계층화하여 설계하고, 세밀하게 표현하여야 하는 요소들은 JSTOR의 저자 형태 구분처럼 속성을 이용하여 추가적인 의미를 부여하는 것이 상호운영 면에서 나을 듯하다. 분류 요소를 사용할 경우에는 반드시 분류체계에 대한 정의가 선행되어야 하며, 권호 표기, 페이지 표기는 2004년 과학기술 정보표준화위원회에서 권고한 표준 표기방식을 준수하고 언어코드의 경우는 ISO639 국제 표준을 따르는 메타데이터를 설계하는 것이 좋을 것이다. 이러한 제안은 메타데이터 상호운영과 통합을 위한 여러 연구에 기초 자료로 사용하는데 도움이 될 것으로 기대한다. 향후 연구로는 본 논문의 자료를 토대로 통합관리와 상호운영에 있어서 반드시 필수적인 요소를 포함하고 내용 규칙과 의미구조를 통일한 통합 스키마의 설계를 할 예정이다. 설계할 통합스키마에는 기술용 메타데이터뿐 만 아니라 원문이나 문헌에 포함된 이미지 등의 연계를 위하여 구조용 메타데이터도 포함할 계획이다.

참 고 문 헌

[1] http://www.ontotext.com/research/cubist

[2] Priscilla Caplan, 오동근 역, *메타데이터의 이해,* 태일사, 2004.

[3] 이경남, "전자기록의 장기적 보존을 위한 보존메타데이터 요소 분석", 기록학연구, 제14권, pp.191-240, 2006.

[4] 이혜진, 송인석, "효율적 정보자원 공유를 위한 서지 메타데이터 XML DTD 개발", 한국콘텐츠학회 종합학술대회 논문집, 제2권, 제2호, pp.427-433, 2004.

[5] 유사라, "메타데이터 주제 국내 연구동향 분석", 한국문헌정보학회지, 제44권, 제2호, pp.405-426, 2010.

[6] http://www.wiley.com

[7] http://www.emeraldinsight.com

[8] http://www.iop.org

[9] http://www.jstor.org

[10] http://www.karger.com

[11] http://www.natureasiapacific.com

[12] http://www.elsevier.com

[13] http://www.springer.com

[14] http://www.chemistry.org/portal/a/c/s/1/home.html

[15] http://www.bione.org

[16] A. Rosenthal, Leonard J. Seligman, "Data Integration in the Large: The Challenge of Reuse," Proceeding of the 20th VLDB Conference, pp.669-675, 1994.

[17] Margaret St. Pierre, *Issues in Crosswalking Content Metadata Standards*, NISO White

Paper, 1998.
[18] E. Rahm and P. A. Bernstein, "A survey of approached to automatic schema matching," The VLDB Journal, Vol.10, pp.334-350, 2001.
[19] http://www.ontoframe.kr/S3/main.jsp
[20] 과학기술정보표준화위원회, *과학기술잡지 권/호 패턴 표준*, 2004.

저 자 소 개

이 민 호(Min-Ho Lee) 정회원

- 2000년 : 충남대학교 대학원 컴퓨터공학과 졸업(석사)
- 2006년 : 충남대학교 대학원 컴퓨터공학과(박사수료)
- 2000년 ~ 2001년 : 데이콤 중앙연구소 연구원
- 2001년 ~ 현재 : 한국과학기술정보연구원 정보기술연구실 선임연구원

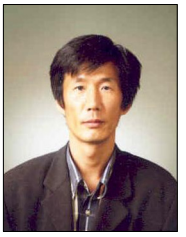
<관심분야> : 정보검색 및 추출, 정보보호, 분산시스템

이 원 구(Won-Goo Lee) 정회원

- 2000년 : 한남대학교 대학원 컴퓨터공학과 졸업 (공학석사)
- 2005년 : 한남대학교 대학원 컴퓨터공학과 졸업 (공학박사)
- 2005년 ~ 현재 : 한국과학기술정보연구원 정보기술연구실 선임연구원

<관심분야> : 데이터베이스, 지식관리, 과학데이터

윤 화 묵(Hwa-Mook Yoon) 정회원



- 1992년 : 서울산업대학교 전자계산학과 졸업(학사)
- 1997년 : 공주대학교 대학원 전자계산학과 졸업(석사)
- 2008년 : 배재대학교 컴퓨터공학과 졸업(박사)
- 현재 : 한국과학기술정보연구원 정보기술연구실 책임연구원

<관심분야> : 데이터베이스, 정보검색, 온톨로지

신 성 호(Sung-Ho Shin) 정회원

- 2000년 : 경북대학교 경영학과 졸업(학사)
- 2002년 : 경북대학교 대학원 경영학과 졸업(석사)
- 2008년 : 배재대학교 컴퓨터공학과 졸업(박사)
- 2002년 ~ 현재 : 한국과학기술정보연구원 정보기술연구실 선임연구원

<관심분야> : 데이터통합, 데이터품질, IS평가

류 재 철(Jae-Cheol Ryou) 정회원

- 1988년 : Iowa State University 전산학과 졸업 (석사)
- 1990년 : Northwestern University 전산학과 졸업 (박사)
- 1991년 ~ 현재 : 충남대학교 전기정보통신공학부 교수
- 2003년 ~ 현재 : 인터넷침해대응기술연구센터장

<관심분야> : 정보보호, 네트워크보안, 암호학, 보안 프로토콜