

# 생명과학 Open Archiving Community를 위한 메타데이터 스키마 설계

안부영

한국과학기술정보연구원 바이오인포매틱스센터  
ahnyoung@kisti.re.kr

## Design of Metadata Schema for Biology & Bioinformatics Open Archiving Community

Bu-young Ahn

CCBB (Center for Computational Biology & Bioinformatics), KISTI

### 요 약

한국과학기술정보연구원(KISTI) 바이오인포매틱스센터(CCBB)에서는 생명과학 관련 주제별 Open Archiving Community의 구성과 운영을 통한 연구자들 간의 정보교환을 유도하고, 더불어 논문뿐만 아니라 세미나, 연구노트 등의 최신의 연구 정보를 공유할 수 있도록 생명과학 Open Archiving Community 시스템을 구축하여 운영하려고 한다. 본 community에서 수집, 구축, 서비스할 정보의 종류는 아티클, 학위논문, 연구 보고서, 발표자료, 연구노트, 실험데이터, 전자자료 등이다.

그러므로, 전 세계적으로 많이 활용되고 있는 Dublin-core, Marc21, MODS를 비교 분석하여 MODS 메타데이터를 기준으로 아티클과 같은 문헌정보 뿐만 아니라 실험결과, 연구노트 등의 비문헌정보도 유연성 있게 적용할 수 있는 생명과학 관련 주제별 Open Archiving Community를 위한 메타데이터 스키마를 설계하였다.

### 1. 서 론

우리나라는 인간 유전체사업에 참여한 다른 나라들보다 생명과학 분야의 연구가 뒤떨어져 있지만 현재 IT, BT 관련 분야에서 많은 산학연 연구자들이 연구에 매진하고 있다. 연구자들이 연구를 하기 위해서는 생명정보 데이터베이스와 그 데이터를 분석할 수 있는 분석도구 및 그 동안의 연구결과물이 집적되어 있는 문헌 및 연구결과 정보가 필요하다.

한국과학기술정보연구원(KISTI) 바이오인포매틱스센터(CCBB)에서는 연구자들이 필요로 하는 생명정보 데이터베이스와 분석도구를 홈페이지를 통해 서비스하고 있으나, 바이오인포매틱스 관련 문헌 및 연구결과 정보 서비스는 제공하고 있지 않다.

이에, 바이오인포매틱스센터 홈페이지를 이용하는 이용자들이 각자의 연구결과와 논문, 세미나 자료 등을 공유하며 교환할 수 있는 생명과학 관련 주제별 Open Archiving Community의 구성과 운영을 위한 시스템을 구현하여 제공하고자 한다.

Open Archiving Community가 구현되어 운영된다면 연구자들 간의 정보교환과 더불어, 논문뿐만 아니라 세미나 발표자료, 연구노트, 실험데이터 등 최신의 연구정보가 공유될 수 있을 것이다.

본 시스템에서 유통되는 정보는 메타데이터이며 아티클, 보고서, 학교교재, 세미나자료, 전자자료 등을 원문과 url까지 포함하여 메타데이터 레지스트리를 구축하고 제공하여야 하므로 유연성있는 메타데이터 스키마를 설계하여 시스템에 적용하였다.

### 2. 메타데이터 스키마 설계

#### 2.1 메타데이터 필드별 비교

메타데이터로서 역할을 하려면 다음과 같은 기본조건을 만족해야 한다. 본 시스템에서는 메타데이터로서 출현한 지는 얼마되지 않았지만 다양한 정보와 웹사이트에 수록된 공개된 자료와 전자매체까지도 표현이 가능한 MODS(Metadata Object Description Schema)와 MARC을 기본으로 스키마를 구성하였다. 현재 MODS를 활용한 시스템 사례가 부족하다는 것과 MARC의 필드들이 복잡하다는 문제점은 있지만, 설계된 스키마를 적용한 프로토타입 시스템을 구현하여 운영하면서 문제점이 발생할 경우 개선해 나갈 여지는 충분하다.

<표 1> 메타데이터 기본조건

조건	내용
Descriptive(기술사항)	서명, 저자, 초록, 주제명 등
Technical & Structural(기술 및 구조정보)	파일크기, 파일유형, 필요한 소프트웨어, 표현 방법(presentation instruction) 등
Administrative(관리정보)	메타데이터의 메타 레코드번호, 레코드 날짜, 레코드 작성기관 등
Rights(지적소유권)	저작권 및 사용권 등
Management(운영정보)	가격, 이용제한사항 등(일반적으로 소유기관에 의해 이루어짐)

문헌정보 네트워크의 데이터베이스를 설계하기 위하여 MODS, DC(Dublin Core), MARC를 필드별로 비교하여 보았다. Open Archiving Community에서는 저널, 아티

클, 전자자료, 세미나 자료 등의 다양한 정보를 공유하고 검색하여야 하기에 메타데이터 스키마는 복잡해질 수 있다. 그러나 본 시스템은 정보기술 활용이 미흡한 생명과학 연구자들이 주로 이용할 것을 감안하여 누구라도 용이한 접근이 가능하도록 스키마를 단순화하였다.

<표 2> MODS와 DC 요소 비교

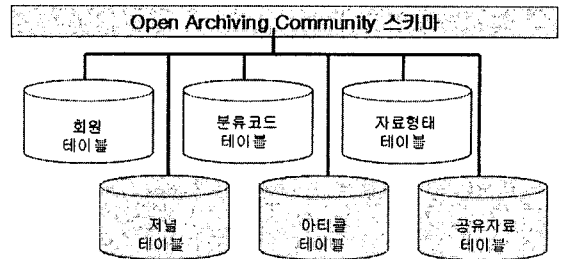
MODS	한글명	DC	한글명
titleInfo	서명사항	Title	표제
name	저자사항	Creator Contributor	제작자 및 기타기여자
typeOfResource	자료유형	Type	유형
genre	장르		
originInfo	발행사항	Publisher Date	발행처 날짜
language	언어	Language	언어
physicalDescription	형태사항	Format	표현형식
abstract	초록	Description	내용기술
tableOfContents	목차	Description	내용기술
targetAudience	이용대상	Audience	이용대상자
note	주제	Description	내용기술
subject	주제	Subject	주제
classification	분류	Subject	주제
relatedItem	관련정보	Relation	관계
identifier	식별기호	Identifier	식별기호
location	소장위치		
accessCondition	이용제한	Rights	권리관리
extension	로컬정보		
recordInfo	레코드정보		

<표 3> MODS와 MARC 태그 비교

MODS	한글명	MARC 필드
titleInfo	서명사항	245 \$a, \$b, \$n, \$p 등
name	저자사항	100, 700 \$a, \$u, 110, 710 \$a
typeOfResource	자료유형	006/00 전자자료 008 leader부의 22, 23 column
genre	장르	655 \$a
originInfo	발행사항	250 \$a, 260 \$a,\$b, \$c, 265 \$a 773 \$t, \$a, \$b, \$d, \$g, \$k, \$p, \$r, \$x 등
language	언어	008 leader부의 35-37 column
physicalDescription	형태사항	300 \$a, \$b, \$c 등
abstract	초록	520 \$a
tableOfContents	목차	505 \$a
targetAudience	이용대상	521 \$a
note	주제	500 \$a, 502 \$a, 506 \$a, 510 \$a, \$c, 513 \$a, \$b 등
subject	주제	650 \$a, 653 \$a
classification	분류	050 \$a, 080 \$a
relatedItem	관련정보	440 \$a, 490 \$a 등
identifier	식별기호	538 \$a, 856 \$a
location	소장위치	049 \$a, \$i 등
accessCondition	이용제한	506 \$a
extension	로컬정보	900 \$a
recordInfo	레코드정보	040 \$a, \$b, \$c 등

2.2 자료종류별 메타데이터 스키마 설계

Open Archiving Community의 데이터를 수집, 관리하기 위하여 메타데이터 레지스트리 스키마를 저널/아티클/공유자료/회원/분류코드/자료형태 테이블 등 6개로 구성하였다.



<그림 1> 스키마 구성도

저널과 아티클의 스키마는 본 시스템에서 서비스되는 정보가 저작권에 관련없이 커뮤니티 회원들이 제공한 자료만을 공유하고 교환한다는 점을 고려하여 설계하였다. 회원들이 공유할 수 있도록 등록된 저널과 아티클의 위치를 알려주는 url을 표기할 수 있는 필드 및 저널을 등록할 수 있도록 제한한 이용자의 id를 표기할 수 있는 필드 등이 이러한 점을 고려한 필드라고 할 수 있다.

<표 4> 저널 스키마

필드명	내용
c_num	자료의 유일성을 보장하는 번호(oaj0000001)
name	저널의 이름을 저장하는 필드
spec_s	저널의 권호시작에 대한 정보를 저장하는 필드
spec_e	저널의 권호끝에 대한 정보를 저장하는 필드
issn	저널의 ISSN 정보를 저장하는 필드
eissn	저널의 E-ISSN 정보를 저장하는 필드
pub_plc	저널을 발행한 지역 및 국가정보
pub_org	저널 발행기관 정보를 저장하는 필드
pub_sdate	저널 발행일 시작 정보를 저장하는 필드
pub_edate	저널 발행일 끝 정보를 저장하는 필드
lang	저널자료의 언어형태 정보를 저장하는 필드
keywords	저널검색을 위한 관련 키워드 저장하는 필드
g_code	분야별 분류를 위한 코드값을 저장하는 필드
url_addr	해당저널 URL 주소를 저장하는 필드
r_gendate	저널 등록 날짜정보를 저장하는 필드
r_moddate	저널 수정 날짜정보를 저장하는 필드
r_generator	저널 등록을 한사람의 이름 또는 아이디 필드

요즘 들어 확산되고 있는 Open Access 기반 저널과 아티클의 메타데이터만을 본 시스템에서 설계한 스키마에 맞도록 레지스트리에 저장한다. 이용자들이 키워드 또는 분류코드 등으로 검색을 수행하면 서지정보(간략보기/상세보기)를 제공받게 되며, 원문을 보고자 하면 서지정보에 포함된 제목 또는 url 필드를 클릭하면 원자료가 있는 웹사이트로 이동하게 된다.

<표 5> 아티클 스키마

필드명	내용
c_num	자료의 유일성을 보장하는 번호(0aa0000001)
name	아티클의 이름을 저장하는 필드
author	아티클 저자 정보를 저장하는 필드
sub_author	아티클 공동저자 정보를 저장하는 필드
j_name	아티클이 수록된 저널명 저장하는 필드
j_spec	수록저널 권호에 대한 정보를 저장하는 필드
pub_plc	수록저널을 발행한 지역 및 국가정보
pub_org	수록저널 발행기관 정보를 저장하는 필드
pub_date	수록저널 발행일 정보를 저장하는 필드
lang	아티클자료의 언어형태 정보를 저장하는 필드
summary	아티클의 요약정보를 저장하는 필드
keywords	저널검색을 위한 관련 키워드 저장하는 필드
g_code	분야별 분류를 위한 코드값을 저장하는 필드
url_addr	해당 아티클 URL 주소를 저장하는 필드
r_gendate	아티클 등록 날짜정보를 저장하는 필드
r_moddate	아티클 수정 날짜정보를 저장하는 필드
r_generator	아티클 등록을 한사람의 이름 또는 아이디 필드

공유자료는 본 시스템을 이용하는 연구자간 공유가 가능한 정보로써 회원들이 자발적으로 등록한 자료이다. 공유자료의 종류는 본 시스템의 요구분석을 위한 이용자 조사에서 응답자들이 모든 자료를 이용했으면 한다는 의견을 주었다. 이에 발행전후의 아티클, 학위논문, 연구보고서, 학교교재(대출출판물 등), 세미나 또는 학술대회에서의 발표자료, 전자자료, 기타자료 등의 모든 자료를 포함 할 수 있도록 설계하였다. 전자 자료 등을 등록 할 수 있도록 url을 표기하는 요소도 지정하였다.

<표 6> 공유자료 스키마

필드명	내용
c_num	공유자료의 유일성을 보장하는 번호(arc0000001)
name	공유자료의 이름을 저장하는 필드
c_type	공유자료의 유형정보를 저장하는 필드
issn_isbn	자료의 ISSN 또는 ISBN 정보를 저장하는 필드
summary	자료의 요약정보를 저장하는 필드
pub_date	자료의 발행 또는 등록일 정보를 저장하는 필드
lang	자료의 언어형태 정보를 저장하는 필드
keywords	자료검색을 위한 관련 키워드 저장하는 필드
g_code	분야별 분류를 위한 코드값을 저장하는 필드
url_addr	해당저널 URL 주소를 저장하는 필드
r_gendate	공유자료 등록 날짜정보를 저장하는 필드
r_moddate	공유자료 수정 날짜정보를 저장하는 필드
r_generator	자료를 등록한 사람의 이름 또는 아이디 필드

4. 결 론

외국의 사례를 보면 arXiv(물리학, 수학, 컴퓨터, 천문학), Cogprint(심리학, 신경과학), RePEc(경제학) 등 연구자들이 주도가 되어 운영되는 주제기반, 기관중심 아카이빙 커뮤니티가 구축되어 활발히 이용되고 있지만 우리나라에는 아직 이런 커뮤니티가 운영되고 있지 않다. 또한 외국에서는 아카이브를 구축할 수 있는 소프트웨어인 E-Prints 소프트웨어의 보급과 함께 기관중심의 아카

이브가 구축되어 있으나, 국내의 경우에는 E-print 아카이브가 구축되어 있지 않을 뿐만 아니라 양질의 정보를 제공하고 있는 외국의 E-print 아카이브 관련기관에 대한 정보도 부족한 실정이다. 현재까지 Open Access 기반 문헌자료의 유통 자체가 활발하지 않은 국내 현실을 반영한 결과라고 볼 수 있을 것이다.

이런 상황에서 생명과학 Open Archiving Community의 구성 및 운영은 시기적절할 뿐 아니라, 정보자원의 공유·교환이 가능하고, 가상의 연구·학술 공간을 제공해 줌으로써 연구자들의 연구효율을 높이는 데 크게 기여할 것이라고 기대되는 바이다.

Open Archiving Community가 원만히 운영되기 위해서는 시스템적인 지원도 필요하지만 먼저 Open Archiving Community 활동에 참여할 수 있는 국내 생명과학 관련연구자와 관련기관을 중심으로 홍보를 실시하여 자발적이고 능동적인 참여를 유도해야 할 것이다. 뿐만 아니라, Open Archiving Community에 많은 연구자들이 참여할 수 있도록 정보공유정책을 수립하여 적용할 수 있어야 한다. 생명과학 Open Archiving Community가 성공적으로 운영된다면 다른 학문분야로의 확대 시행도 가능하리라 예상된다.

참고문헌

- [1] 이수상, MODS(Metadata Object Description Schema - 디지털도서관의 서지정보 표준 메타데이터, 2004년도 KISTI "메타데이터 설계 및 적용사례" 세미나 발표자료, 2004, 7.16
- [2] 이창열, ONIX 메타데이터 체계와 구현, 2004년도 KISTI "메타데이터 설계 및 적용사례" 세미나 발표자료, 2004, 7.16
- [3] Arun Jagota, eDICTIONARY OF BIOINFORMATICS, Bioinformatics by the Bay Press, 2001
- [4] 이수상, 학술 정보 유통에 OAI 적용에 관한 연구, KERIS, 2004.6.
- [5] 정원규, 철학 텍스트들의 내용분석에 의거한 디지털 지식자원 구축을 위한 기초적 연구, 서울대학교 철학사상연구소, 2002
- [6] 설성수, 송충한, 기초과학연구의 분야분류체계 개발 연구, 한국과학재단 연구보고서, 1999.4.
- [7] 김상준, E-print 아카이브를 위한 OAI 응용 디지털 도서관, 충남대 문헌정보학과 Term Paper, 2004.6.3.
- [8] 황해경, 김혜선, 최선희, 오픈액세스기반 지식정보 저장소 구축에 관한 연구, 비블리아학회, 2005.5.21.
- [9] 윤정옥, MARC의 구조와 기능, KISTI 국내정보실 세미나 자료, 2004.7.
- [10] <http://www.doaj.org>
- [11] <http://www.lub.lu.se/ncsc2002>
- [12] <http://www.loc.gov/standards/mods>
- [13] <http://society.kisti.re.kr>
- [14] <http://www.koreamed.org>
- [15] <http://www.pubmed.org>