생물다양성 데이터교환을 위한 메타데이터 스키마 설계

안부영, 조희형, 안성수, 박형선
한국과학기술정보연구원 바이오인포매틱스센터
{ahnyoung, choh2, ssahn, seonpark}@kisti.re.kr

Design of Metadata Schema for Biodiversity Data Exchange

Bu-young Ahn, Hee-hyung Cho, Sung-soo Ahn, Hyung-seon Park
CCBB (Center for Computational Biology & Bioinformatics), KISTI

요 약

생물다양성은 육상 생태계, 해양과 기타 수생 생태계와 이들의 복합 생태계를 포함하는 모든 원천에서 발생한 생물체의 다양성을 말하며, 종내·종간 및 생태계의 다양성을 포함한다. 지구상에 존재하는 생물이 매우 다양하듯이 생물다양성을 표현하는 데이터 또한 매우 다양하게 사용되고 있다. 본 논문에서는 먼저 생물다양성 데이터의 정보공유 및 교환을 위해 생물다양성 관련 국제기구에서 제안된 데이터 표준 및 데이터 교환 프로토콜을 알아보고, 이러한 데이터 표준과 프로토콜을 기반으로 국내 생물다양성 데이터 공유 및 교환을 위한 생물다양성 메타데이터 스키마를 크게 생물종 정보와 종정보에 관한 참조(reference) 정보로 나누어 설계하여 제시하고자 한다.

1. 서 론

생물다양성은 육상생태계, 해양과 기타 수생생태계와 이들의 복합생태계를 포함하는 모든 원천에서 발생한 생물체의 다양성을 말하며, 종내·종간 및 생태계의 다양성을 포함한다.

<표 1> 한국에서 밝혀진 생물종 수

대분류군		소분류군		종 수	
동 물 18,029	척추동물 1,440	포유류		100	
		어 류		905	
		양서·파충류		41	
		조 류		394	
	무척추동물 3,564	해면	204	자포	224
		편형	123	윤형	159
		구두	1	내항	1
		태형	145	완족	9
		성구	9	연체	997
		환형	380	완보	49
		절지	1,028	모악	39
		극피	107	미색	89
	13,025	곤충		11,853	
		거미		1,172	
식 물 8,271	고등식물 4,662	단자엽식물		842	
		쌍자엽식물		2,815	
		양치·나자식물		314	
		선태류		691	
	하등식물 3,609	규조류		1,512	
		편모조류		316	
		담수녹조류		1,064	
		윤조류		27	
		해조류		690	
균류(지의류포함)		3,528		1,625	
원생생물				736	
원핵생물				1,167	
총 29,828종					

지구상에는 약 170만종의 생물종이 알려져 있으며, 조사되지 않은 생물종을 감안할 경우 약 1,250만종(범위: 5백만종-1억종)이 살고 있는 것으로 추정되고 있다(Global Environment Outlook 2000, UNEP). 국내에는 10만종 이상의 생물종이 서식하고 있는 것으로 추정되며 현재까지 조사된 생물종 수는 <표 1>에서 보듯이 29,828종이다(국내 생물종 조사, 1996, 환경부).

이러한 생물다양성 데이터는 이 데이터를 활용하고자 하는 연구자들에게 없어서는 안될 필수적인 데이터이다. 그런데, 생물다양성 데이터는 각자의 특성이 다양하기에 그런 특성만을 중시하여 각 전문가들에 의해 데이터베이스가 구축되고 있다. 생물다양성 각자의 특성도 중요하지만 이 데이터를 활용하기 위해서는 데이터 공유와 교환을 위한 메타데이터 표준이 매우 절실한 실정이다.

이에, 본 논문에서는 먼저 생물다양성 관련 데이터 표준 및 데이터 교환 프로토콜을 알아보고, 이러한 데이터 표준과 프로토콜을 기반으로 완벽한 데이터의 표준을 제정하기는 어렵지만, 국내의 생물다양성 데이터를 보다 효율적으로 교환 및 공유할 있는 생물다양성 메타데이터 항목을 추출하여 스키마를 설계하고자 한다.

2. 생물다양성 데이터 표준현황

2.1. DawinCore

많은 생물다양성 데이터는 전세계적으로 각각의 특성에 맞게 정보 콘텐트, 스키마, 구조 등이 결정되어 데이터베이스에 저장되어 서비스되고 있다. 이렇게 저장된 데이터베이스는 일련의 관련성이 있기에 이를 이용하여 생물다양성 데이터를 검색하고 데이터베이스에 접근할 수 있다. 이렇게 생물개체의 공통된 특성을 추출하여 생물다양성 데이터로의 접근을 쉽게 할 수 있도록 만든 데이터 형식이 DarwinCore 형식이다.

DarwinCore 형식은 XML 스키마로 정의되고 총 48개의 Element로 구성되어 있으며, 이 중 5개(DateLastModified, InstitutionCode, CollectionCode, CatalogNumber,

ScientificName)는 필수항목이고 나머지는 선택항목이다.

<표 2> DarwinCore 데이터항목

DateLastModified	최종갱신일	MonthCollected	수집월
InstitutionCode	기관코드	DayCollected	수집일
CollectionCode	소장코드	JulianDay	율리우스력
CatalogNumber	목록번호	TimeOfDay	수집시간
ScientificName	학명	ContinentOcean	육상, 해상
BasisOfRecord	개체단위	Country	국가
Kingdom	계	StateProvince	도
Phylum	문	County	시
Class	강	Locality	지역명
Order	목	Longitude	경도
Family	과	Latitude	위도
Genus	속	CoordinatePrecision	좌표정밀도
species	종	BoundingBox	좌표범위
Subspecies	아종	MinimumElevation	최저고도
ScientificNameAuthor	명명자	MaximumElevation	최고고도
IdentifiedBy	동정자	MinimumDepth	최저깊이
YearIdentified	동정년	MaximumDepth	최대깊이
MonthIdentified	동정월	Sex	성별
DayIdentified	동정일	PreparationType	조직표본유형
TypeStatus	학명유형	IndividualCount	개체수
CollectorNumber	수집번호	PreviousCatalogNumber	이전목록번호
FieldNumber	수집지역번호	RelationshipType	관계유형
Collector	수집자	RelatedCatalogItem	관련목록항목
YearCollected	수집년	Notes	주석

DarwinCore는 현재 GBIF, The Species Analyst, MaNIS(Mammal Networked Information System), OBIS(Ocean Biogeographic Information System) 프로젝트 등에서 원래의 형태 또는 확장된 형태로 사용되고 있다.

2.2. ABCD (Access to Biological Collection Data)

DarwinCore는 분산된 생물다양성 데이터제공자의 데이터를 통합할 수 있도록 여러 프로젝트에 효과적인 기반을 제공하였다. 하지만 DarwinCore에서 현재 제공하고 있는 항목은 동물학 데이터에 편중된 특징이 있어 박테리아, 균류, 식물, 원핵생물, 바이러스 등의 데이터 범위를 포함하는 데는 한계가 있다. 이러한 문제를 해결하기 위하여 TDWG(Taxonomic Databases Tasking Group)에서 2000년부터 ABCD 스키마가 개발되었고 관련 프로젝트에서 새로운 데이터 표준으로 자리매김하고 있는 상태이다.

ABCD 스키마는 DarwinCore보다 복잡하고 계층적이며, OriginalSource, DatasetDerivations, Units 항목으로 구분된다. 버전 1.2는 총 371개의 항목을 가지고 있고, 필수항목으로는 SourceInstitutionCode, SourceName, SourceLastUpdatedDate, DateSupplied, UnitID 등 5개이고 나머지는 선택항목이다.

현재 700여개의 항목을 가지고 있는 버전 2.05는 TDWG에 제출되어 스키마에 대한 의견수렴을 진행하고 있다.

2.3. DiGIR (Distribute Generic Information Retrieval)

DiGIR는 XML 스키마를 따르는 분산된 생물다양성데이터와 데이터베이스의 정보를 검색하기 위한 클라이언트-서버 프로토콜로 HTTP상에서 XML 메시지를 주고받는데 사용된다. DiGIR 프로토콜은 CODATA Biological Collection Data 그룹과 Taxonomic Data Working Group에서 정의하였고 현재 DarwinCore 형식의 데이터를 교환할 때 사용할 수 있도록 소프트웨어 패키지가 PHP, Java로 개발되어 오픈 소스로 제공되고 있다. DiGIR 프로토콜에는 3가지의 메시지 타입(Metadata, Inventory, Search)이 존재한다.

2.4. BioCASE (A Biological Collection Access Service for Europe)

BioCASE 프로젝트는 유럽위원회(the European Commission)의 지원을 받아 2001년 11월에 시작된 프로젝트이다. 이 프로젝트의 목적은 유럽의 연구자들에게 웹 기반 생물다양성 정보서비스를 제공하고 구현하도록 하는 것이다. 현재 유럽의 30개국과 이스라엘의 35개 기관이 참여하고 있으며, 각 기관은 생물다양성 데이터에 대한 국가 노드(데이터제공자) 역할을 수행하면서 데이터수집과 공유에 협력하고 있다.

BioCASE 프로토콜은 BioCASE 프로젝트에 사용된 것으로 데이터베이스에 대한 질의와 응답방법 등을 정의하고 있으며, 데이터 교환 형식으로는 ABCD 스키마를 사용하고 있다. BioCASE 프로토콜은 3개의 request 방법(search, scan, capabilities)을 정의하고 있다.

3. 생물다양성 메타데이터 스키마 설계

2장에서는 DarwinCore 데이터 형식과 ABCD 스키마 형식을 살펴보았다. 이를 참고하여 3장에서는 DarwinCore 형식을 확장하여 국내에서 활용 가능한 생물다양성 메타데이터 스키마를 설계하고자 한다.

3.1. 기존 생물다양성 데이터항목 분석

<표 3> 기구축된 생물다양성 DB 항목

담수어류		고등균류		자생식물종자	
num	과별고유번호	cname	한국보통명	an	일련번호
family	과 이름	s_name	학명	kname	한국명
branch	과 현황	s_id	종명	ename	과명
detail	과 설명	g_id	속명	boname	학명
id	고유번호	f_id	과명	s_name	종명
name	이름	da_id	분포지역	s_id	종명_id
family	과명	use	이용별	g_id	속명_id
genus	속명	habitat	서식지별	f_id	과명_id
species	종명	ecotype	군락형태별	assortment	구분/분류
eng_name	영명	picture	관련사진	habitat	서식지
dialect	방언	spore	포자	ecotype	생태형
length	전장	bk_id	수록버섯도감	picture	사진
branch	분류	con	내용	kcon	한글특징
speciality	특징			econ	영문특징
life	생태				
distribution	분포				
reference	참고사항				
area	지역				
river	수계				

국내에서 구축된 생물다양성 데이터베이스는 수십종에 이른다. 그중에서 KISTI 생물다양성 데이터베이스 구축 사업으로 생산된 DB 일부의 항목을 <표 3>과 같이 분석해 보았다. <표 3>에서 보는 것과 같이 생물종이 다양한 만큼 다양한 데이터 항목으로 구성된 것을 볼 수 있었다.

3.2. 메타데이터항목 설계

현재 국제생물다양성정보기구 (GBIF: Grobal Biodiversity Information Facility)를 중심으로 Unified Biodiversity Information Protocol(UBIF)에 대한 제안서가 제출되었고 이것에 대한 논의가 활발히 진행되고 있다. 이에 본 논문에서는 향후 국제적인 교환을 고려하여 위에서 분석된 기존의 데이터항목과 DarwinCore 데이터 형식을 비교하면서 국내 생물다양성 데이터의 교환과 공유를 위한 생물다양성 데이터항목을 <표 4>와 같이 구성하였다.

<표 4> 생물다양성 메타데이터항목

종(species) 정보		참고(reference) 정보		
계(kingdom)	kingdom subkingdom	종설명 정보	설명 정보	종설명 특이사항
문(division)	division phylum subphylum subdivision		특징 정보	생물학적특성 번식방법 성분 용도 수명
강(class)	superclass class subclass intraclass	멀티미디어 정보	이미지	사진 그림 정보제공자
목(order)	superorder order suborder intraorder		영상	소리 동영상 정보제공자
과(family)	superfamily family subfamily intrafamily	서식 정보	분포 정보	위경도 주소
속(genus)	tribe subtribe genus scetion subgenus		형태 정보	수상 육상
종(species)	species subspecies	참고 문헌 정보	단행본	서적명 저자 출판사 발행년도
일반명 (common name)	한글명 영어명 지역명 북한명 어원 동의이명 기타명		논문	저널명 권호사항 제목 저자 페이지 년도 언어
			소장처	DB 웹사이트 기관명
명명자	명명자이름 소속기관 이메일 전화 주소 사진	관리 정보	관련 정보	채집사항 파생정보 국제협력 기원 상태정보
제어번호	고유 데이터 번호		소유권 정보	입력일 저작권 정보레벨

<표 4>의 데이터항목에서 보는 것과 같이 크게 종(species) 정보와 참고(reference) 정보를 나누어 구성하였으며 참고정보에는 관리정보 항목을 두어 데이터의 생성과 정보공개를 위한 레벨정보까지를 총체적으로 기술할 수 있도록 하였다.

4. 결론

국내에 기 구축된 생물다양성 데이터베이스가 제공하는 데이터는 질의형식, 데이터모델, 스키마구조, 사용시스템에서 이질적인 특성이 나타난다. 이는 생물다양성 정보자원의 통합을 어렵게 하는 요인으로 작용하게 된다. 또한 분산된 환경에서 다양한 데이터 형태의 정보를 물리적으로 통합하려면 여러 가지 어려움에 봉착하게 된다. 이런 어려운 문제들을 해결해 보고자 지금까지 미국과 유럽을 중심으로 널리 사용되고 있는 생물다양성 데이터 표준(DarwinCore, ABCD)과 데이터 교환 프로토콜(DiGIR, BioCASE)이 연구되고 있는 것이다.

그러나 외국의 것을 국내에 그대로 적용하면 국내정보가 누락되는 경우가 발생할 수 있다. 그래서 기 구축된 데이터베이스의 스키마를 분석한 후 가장 기본적이고도 필요한 데이터항목을 선별한 후 DarwinCore 스키마를 확장하여 메타데이터 항목을 정의하고 설계하였다.

이렇게 설계된 데이터항목은 생물다양성 정보의 핵심적인 사항만을 표현하기에, 국내 생물다양성 정보의 리포지토리를 구축하는데 활용할 수 있다. 이 리포지토리를 통하여 데이터의 실제 위치를 찾아가도록 연결고리를 제공할 수 있을 것이다.

또한 지속적인 생물 종의 발굴과 관련 연구에 따른 생물종의 데이터, 사진, 연구노트 등이 생물학자에 의해 기록되고 있기에 기록 시점에서부터 본 데이터 항목을 사용한다면 이런 생물다양성 데이터가 실물과 연계되면서 정보공유 및 교환에 큰 역할을 감당하게 될 것이다.

그리하여 국내에서 구축된 생물다양성 데이터베이스에 대한 용이한 접근이 가능하며, 표준화된 데이터베이스 구축, 정보의 고유성 및 정보제공으로 인한 손실을 최소화할 수 있을 것이다. 또한 국제표준을 참고하였기에 국제적인 생물다양성 데이터베이스 공유 및 교환 협력도 가능할 것으로 기대한다.

참고문헌

[1] Global Environment Outlook 2000, UNEP
[2] 안부영, "KISTI 생물다양성 DB 구축현황", 지식정보인프라, 통권 10호, 26-39, 2002.7
[3] 안성수, 박형선, 안부영, 조희형, "생물다양성데이터 검색포탈 구축", 한국콘텐츠학회 2005 춘계종합학술대회 논문집, 124-128, 2005.5.20-21, 동의대학교
[4] 양진호, 이계준, 안부영, 박형선, "XML기반 생물자원정보 관리시스템 설계", 2002년 정보과학회 추계학술대회 논문집, 289-291, 2002.10.25-26, 수원대학교
[5] http://digir.net/schema/conceptual/darwin/2003/1.0/darwin2.xsd
[6] http://bgbm3.bgbm.fu-berlin.de/TDWG/CODATA/Schema/default.htm