KBIF 데이터 노드 구축

안성수[O] 양진호 권창혁 박형선
한국과학기술정보연구원 바이오인포매틱스센터
{ssahn[O], spearjin, narrowpath, seonpark}@kisti.re.kr

Construction of KBIF Data Node

Sungsoo Ahn[O] Jinho Yang, Changhyuk Kwon, Hyung-Seon Park
Center for Computational Biology and Bioinformatics, KISTI

요 약

국제생물다양성정보기구(GBIF)는 전세계의 생물다양성데이터베이스를 네트워크로 연결하고 인터넷을 통한 서비스를 제공하여 생물다양성데이터가 자유롭게 널리 이용될 수 있는 임무를 수행하고 있다. 한국에서는 KISTI가 국가중점노드 역할을 수행하면서 생물다양성데이터 보유기관에 데이터노드 구축 관련 기술과 소프트웨어를 보급하고 있고 현재 한국에서는 2개의 데이터 노드가 구축되어 GBIF의 데이터 포털과 연결되어 있다. 본 논문에서는 GBIF의 생물다양성데이터를 교환하기 위해서 필요한 데이터 표준, 프로토콜, 관련 소프트웨어를 소개하고 데이터 노드 구축 방법을 소개하고 생물다양성데이터의 응용 방법에 대해 논의한다.

1. 서론

최근 컴퓨터가 널리 사용되면서 자연사 박물관, 식물원, 동물원 등의 생물다양성데이터를 교환하고 활용하여 교육, 환경 보호, 산업 등에 활용하려는 움직임이 일어나고 있다. 국제적으로는 1999년에 설립된 국제생물다양성정보기구(Global Biodiversity Information Facility, 이하 GBIF)가 전세계의 생물다양성자원을 네트워크로 연결하고 인터넷을 통한 서비스를 제공하여 생물다양성데이터가 자유롭게 널리 이용될 수 있는 임무를 수행하고 있다[1]. 2001년에 GBIF에 가입한 한국은 한국생물다양성정보기구(Korean Biodiversity Information Facility, 이하 KBIF)로 활동하고 있고 한국과학기술정보연구원(이하 KISTI)은 국가중점노드기구의 역할을 맡아 국내의 생물다양성데이터 보유기관에 데이터 표준과 데이터를 유통할 수 있는 소프트웨어를 교육 및 보급하여 데이터의 유통을 촉진시키는 활동을 하고 있다. GBIF는 현재 전 세계의 75개 생물다양성데이터베이스, 약 3천 8백만 건의 데이터를 GBIF 데이터 포털(http://www.gbif.net)를 통하여 제공하고 있으며 한국은 어류 데이터베이스, 한국 균주 데이터베이스 2개를 GBIF 네트워크에 연결하였다. 본론에서는 GBIF의 웹서비시스 아키텍쳐와 데이터 교환 표준, 프로토콜을 소개하고 데이터 노드 구축 방법에 대한 논의한다.

2. 본론

생물다양성데이터를 교환하기 위해서는 공통의 데이터 형식과 이를 네트워크상에서 교환할 수 있는 프로토콜이 필요하다. 국내에는 이러한 데이터 표준을 현재 KISTI와 한국생명공학연구원 생물자원센터(이하 KRIBB)를 중심으로 설계 및 개발 중이다. 유럽에서는 ABCD 표준 데이터 형식과 BioCASE 프로토콜이 사용되고 있고 미국 등에서는 CODATA의 TDWG(Taxonomic Data Working Group)에서 만든 Darwin Core[2] 표준 데이터 형식과 DiGIR 프로토콜[3]이 사용되고 있다. 전 세계의 생물다양성데이터를 자유롭게 이용하고 접근할 수 있는 것을 목적으로 하는 GBIF는 앞에 언급한 DarwinCore/DiGIR, ABCD/BioCASE의 표준을 지원하면서 시스템 아키텍쳐[4]는 웹서비시스(Web Services)를 지향하고 있다. 즉, SOAP 대신 DiGIR, BioCASE 프로토콜상에서 데이터를 교환하고 UDDI 레지스트리를 이용하여 데이터 노드를 홍보하고 검색할 수 있는 체제이다. 그림 1은 GBIF 웹서비시스 스택이다.

현재 GBIF 데이터 포털에 연결되어 있는 대부분의 생물다양성데이터베이스는 DarwinCore XML 스키마와 DiGIR 프로토콜을 사용하고 있고 유럽의 몇몇 데이터베이스는 ABCD XML 스키마와 BioCASE 프로토콜을 사용하고 있다. KISTI에서 구축한 어류 데이터베이스는 DarwinCore와 DiGIR 프로토콜을

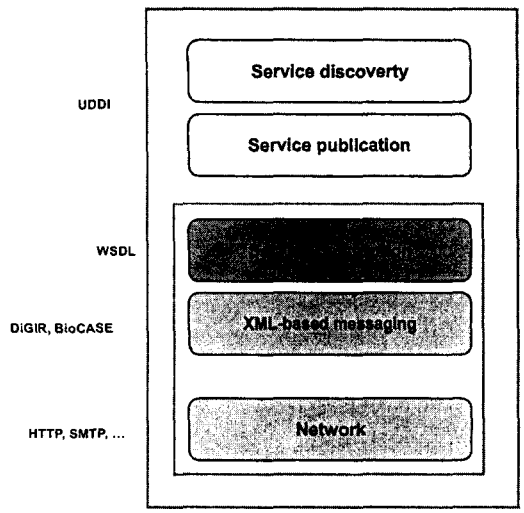

그림 1. GBIF 웹 서비시스 스택

사용하므로 이에 대한 설명을 한다.
DarwinCore는 48개의 항목으로 구성된 XML 스키마로 처음 5개의 항목(DateLastModified, InstitutionCode, CollectionCode, CatalogNumber, ScientificName)이 반듯이 있어야 하고 나머지 항목(BasisOfRecord, Kingdom, ···)은 선택적으로 존재할 수 있다.
DiGIR 프로토콜은 클라이언트가 DarwinCore 스키마 형식을 따르는 분산된 자원(DB, XML 문서)에 대해 질의하고 검색할 수 있게 하고 3개의 메시지 타입(Metadata, Search, Envelope)을 지원한다. 그림 2는 간단한 DiGIR 아키텍쳐이다.

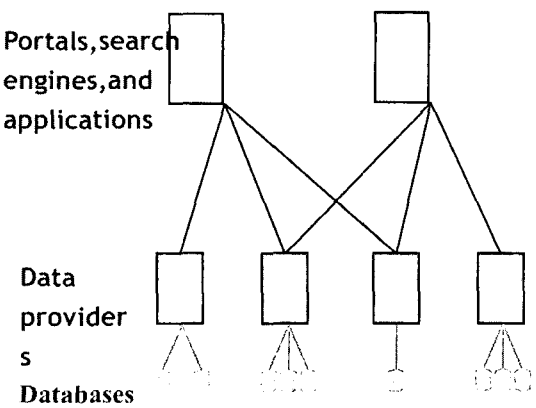

그림 2. DiGIR 아키텍쳐

GBIF 데이터 포털에서 국내 데이터노드의 자원 즉, 생물다양성데이터베이스를 검색할 수 있도록 하기 위해서는 DarwinCore 또는 ABCD 스키마에 맞게 데이터베이스를 구축한 후 이를 서비스하여야 한다. KISTI는 2002년 생물다양성구축사업을 통하여 어류 데이터베이스 외에 19개 데이터베이스를 구축하였고 현재 이를 웹을 통하여 서비스 하고 있다. 그렇지만, 생물다양성 데이터베이스를 구축할 때 데이터 표준 스키마가 존재하지 않아 생물다양성데이터베이스가 각기 다른 테이블 스키마를 가지고 구축이 되었다. KISTI에서는 먼저, 당수 어류 데이터베이스를 DarwinCore XML 스키마 형식의 데이터타입에 적합한 데이터베이스 스키마를 재구성하여 데이터베이스를 재구축 한 후 DiGIR 프로토콜을 사용하는 데이터노드 서버를 구성하여 GBIF 데이터 포털에 연결하였다. 그림 3 참고.

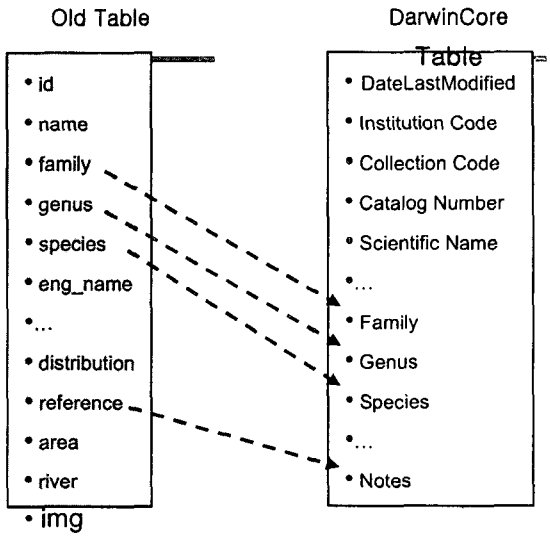

그림 3. DarwinCore 스키마로 변환

또한, 한국생명공학연구원에서도 기존의 균류 데이터베이스를 DarwinCore와 DiGIR 프로토콜을 사용하여 GBIF 데이터 포털에 연결하여 서비스를 하고 있다.

기존의 생물다양성DB가 구축되어진 경우에 이렇게 DarwinCore를 따르는 새로운 테이블 만드는 것이 효과적이다. 그렇지 않고 기존의 테이블을 결합(join)하여 사용할 경우 검색 및 접근에 성능이 저하될 수 있다. 새로운 테이블을 만들 경우 기존의 데이

터베이스와 동기화(sync)하는 것이 문제가 될 수 있다. 데이터베이스의 동기화 및 유지보수 문제는 일주일 또는 한달간 주기적으로 원래 데이블의 데이터를 DarwinCore 형식을 따르는 데이터로 변환하는 프로그램을 만들면 해결 할 수 있을 것으로 생각된다.

KBIF의 국가중점노드 역할을 수행하는 KISTI는 한국생명공학연구원, 국립중앙과학관, 농업생명공학연구원, 산림청 수목원 등의 정부 유관 기관에 GBIF에서 권장하는 DarwinCore 스키마와 DiGIR 프로토콜, 관련 소프트웨어를 교육, 보급하였고 앞으로 더 많은 기관에 관련기술을 보급할 예정이다. 또한 KISTI에 보유한 기존의 생물다양성 데이터베이스를 변환하여 GBIF 데이터 포털에 연결하여 서비스할 예정이다.

GBIF에서는 IT 지식이 없는 현장의 실무자들이 생물다양성데이터를 수집, 저장하고 공유할 때 편리하게 사용할 수 있도록 DataRepository Tool을 제공하고 있다. 이 소프트웨어는 사용자가 엑셀 또는 XML 형식의 데이터를 DarwinCore형식에 맞게 저장한 후 DataRepository에 올리면 이를 데이터베이스와 같은 자원으로 인식하여 데이터를 검색 및 접근 가능하게 하는 기능을 가지고 있다. 이 소프트웨어는 Zope 웹 서버에서 단독으로 사용되거나 GBIF PTK(Portal Tool Kit)과 같이 사용될 수 있다. 현재 KBIF(http://www.kbif.re.kr) 웹 사이트에서는 이를 구축하여 서비스하고 있으며 GBIF에 한국의 새로운 데이터 노드로 등록할 예정이다.

3. 결론

본 논문에서는 GBIF의 활동, 서비스 아키텍쳐, 데이터 표준(DarwinCore), 교환 프로토콜(DiGIR) 그리고 GBIF에서 제공하는 소프트웨어 도구(DataRepository, Portal Tool Kit)에 대해 설명을 하였다. 그리고 기존의 생물다양성 데이터베이스를 어떻게 변환하여 GBIF의 데이터 포털에 연결할 수 있는지에 대해서 설명하였다.

세계 여러 나라의 생물 종(species) 데이터, 종의 표본(specimen) 데이터와 같은 1차 생물다양성데이터네트워크가 구축되고 이러한 데이터를 자유롭게 이용할 수 있게 되면 응용프로그램으로 종 및 표본의 분포도를 표시할 수 있는 프로그램을 만들어 희귀 종을 보호하거나 국가내의 전체 생물 종 분포도를 만들어 환경정책을 결정하는데 활용될 수 있을 것이다. 또한 항만 또는 공항의 세관에서 의심이 가는 동물, 전염병을 옮기는 생물 종의 반입을 통제할 수 있는 프로그램을 만들 수 있고 학생들에게 자연의 소중함을 알리는 교육용 프로그램을 만들 수도 있다.

국제적으로 생물다양성데이터 공유 및 활용의 움직임에 맞추어 국내에서도 생물다양성데이터를 교환할 수 있는 데이터표준을 정립하고 이를 응용할 수 있는 소프트웨어의 개발이 앞으로 향후 과제가 되리라 생각한다. 이러한 과제를 달성하기 위해 생물다양성분야와 컴퓨터공학의 활발한 연구협력[5]이 있어야 하겠다.

[참고문헌]

[1] Research and Societal Benefits of the Global Biodiversity Information Facility, James L. Edwards, BioScience, June 2004, Vol. 54 No. 6

[2] DarwinCore Schema, http://digir.net/schema/conceptual/darwin/2003/1.0/darwin2.xsd (2004년 9월 1일, 접근)

[3] Distributed Generic Information Retrieval (DiGIR), http://digir.net (2004년 9월 1일, 접근)

[4] GBIF Biodiversity Data Architecture, Donald Hobern, GBIF 기술 문서, http://circa.gbif.net/Public/irc/gbif/dadi/library?l=/architecture/gbifbiodiversitydataarch_1/_EN_0.7_ (2004년 9월 1일 접근)

[5] An Introduction to GBIF Biodiversity Informatics, Donald Hobern, GBIF 기술 문서, http://circa.gbif.net/Public/irc/gbif/dadi/library?l=/architecture/gbifbiodiversityinformat_1/_EN_1.0_, (2004년 9월 1일 접근)