

생물다양성데이터 검색포탈 구축

Establishment of Search Portal on Biodiversity Data

안성수, 박형선, 권창혁, 안부영, 조희형
한국과학기술정보연구원

Ahn Sung-Soo, Park Hyung-Seon, Kwon Chang-Hyuk,
Ahn Bu-young, Cho Hee-young
Korea Institute of Science and Technology
Information

요약

본 논문은 국내외의 생물다양성데이터 표준형식과 프로토콜 등을 소개하고 지리적으로 분산된 국내 생물다양성데이터를 통합 검색하여 활용할 수 있는 방법과 국내 생물다양성데이터의 검색포탈을 어떻게 구축하였는지 설명한다. 다음으로 포탈구축에 사용된 데이터 표준, 데이터 교환 프로토콜, 시스템 아키텍처 그리고 소프트웨어 구성요소에 대해 설명하고 검색포탈이 원활히 운영되어지기 위해 생물다양성데이터 소유기관 등에서 필요한 활동과 생물다양성데이터 검색포탈 구축의 결과 및 기대효과 등을 서술한다.

Abstract

This paper introduces data standards and protocols for exchanging biodiversity data, explains ways to construct biodiversity search portal of Korea, and describes software components used to implement the portal. We also discuss the expected activity of data custodians to share the biodiversity data.

I. 서론

최근 컴퓨터와 인터넷이 널리 사용되면서 자연사박물관, 식물원, 동물원, 대학교 실험실 등에서 소장한 생물다양성데이터(표본, 관찰 및 생물 종의 이름 등)를 교환하여 정책결정, 교육, 환경보호, 산업 등에 활용하려는 움직임이 일어나고 있다. 또한 국내외 생물다양성기구에서는 데이터 표준 및 교환 프로토콜을 연구, 개발하고 이를 따르는 소프트웨어를 활용하여 생물다양성인프라 구축하려 하고 있다.

본 논문에서는 먼저 국외에서 널리 사용되고 있는 생물다양성데이터 표준과 데이터교환 프로토콜에 대

해 서술하고 이러한 것을 기반으로 세계의 생물다양성데이터를 자유롭게 범용적으로 이용할 수 있도록 하는 것을 목적으로 하는 국제생물다양성정보기구(Global Biodiversity Information Facility, 이하 GBIF)[1]와 GBIF의 활동, 시스템 아키텍처 등을 소개한다. 그리고 국제적인 프로젝트인 GBIF의 활동에 동참하면서 국내 생물다양성데이터를 네트워크로 연결하고 활용하는 방법을 제안한다. 마지막으로 지리적으로 분산된 생물다양성데이터를 검색하는 검색포탈의 구축 방법, 시스템의 구성요소와 그 기대효과를 서술한다.

II. 생물다양성데이터 표준형식 및 프로토콜

자연사 소장자료(natural history collections)와 관찰데이터 수집물은 어느 특정 시공간에서 한 개체에 대한 자세한 관찰내용을 포함하는 일련의 산물이다.

많은 자연사 소장자료는 현재 부분적으로 디지털화되어 데이터베이스에 저장되고 있고 각 소장자료의 특성에 맞게 데이터베이스 콘텐츠, 스키마, 구조 등이 결정된다. 그렇지만 수많은 소장 자료와 관찰 데이터의 내용에는 일련의 공통성이 존재하고 이를 이용하여 생물다양성데이터를 교환하고 활용할 수 있다. 생물개체의 공통된 특성을 추출하여 생물다양성데이터의 접근을 쉽게 할 수 있도록 만든 형식으로 현재 국외에서 널리 사용되고 있는 데이터 표준형식으로 DarwinCore[2]와 ABCD[3] 스키마가 있다.

국내에서는 한국과학기술정보연구원(이하, KISTI)에서 국내의 표본, 관찰 데이터 등을 표현할 수 있는 데이터 표준(스키마)을 연구하고 있다.

1. DarwinCore와 DiGIR

DarwinCore 는 XML 스키마로 정의되고 총 48개의 항목(Element) 중 5개 항목(DataLastModified, InstitutionCode, CollectionCode, CatalogNumber, ScientificName)은 필수이고 나머지 43개의 항목은 선택적이다.

TDWG(Taxonomic Data Working Group)에서 개발한 DiGIR(Distributed Generic Information Retrieval)[4] 프로토콜은 DarwinCore 형식의 분산된 자원(XML 스키마를 따르는 생물다양성데이터, 생물다양성 데이터베이스)을 검색하기 위한 클라이언트/서버 프로토콜로 3가지의 메시지 타입을 가지고 있다.

- Metadata: 데이터 제공자(Data Provider)와 자원의 메타데이터 정보를 검색한다.

- Inventory: 하나의 개념(학명, 저자, 기관 등)과 관련된 일련의 값을 검색한다.
- Search: 검색 기준에 따라 표본 및 관찰 레코드를 검색한다.

2. ABCD와 BioCASE

DarwinCore에 기반한 ABCD스키마는 동물학, 박테리아, 균류, 식물, 원핵생물, 바이러스 등의 광범위한 생물다양성데이터를 수용할 수 있게 현재 개발되고 있다. ABCD 스키마는 크게 OriginalSource, DatasetDerivations, Units 항목으로 구분되고 버전 1.2는 총 371 개의 항목을 가지고 있다. 이 스키마는 DarwinCore보다 복잡하고 계층적이지만 필수항목은 SourceInstitutionCode, SourceName, SourceLastUpdatedDate, DateSupplied, UnitID 5가지이고 그 외의 항목은 선택적이다.

BioCASE[5] 프로토콜은 BioCASE 프로젝트에 사용된 것으로 데이터베이스 대한 질의와 응답 방법을 정의하고 데이터 교환 형식으로는 ABCD 스키마를 사용하고 있다. BioCASE 프로토콜은 현재 3개의 요청(request) 방법을 정의하고 있다.

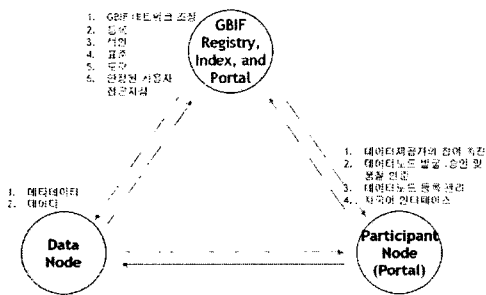
- search: 데이터베이스를 검색하는 방법을 정의하고 실제적으로는 SQL SELECT 문을 XML 형식으로 변환한다.
- scan: xpath로 참조되는 하나의 개념에 대한 요청으로 SQL에서 SELECT DISTINCT문을 나타낸다.
- capabilities: 데이터 제공자의 데이터베이스에 어떠한 개념들이 정의되어 있는지 요청하는 것이다.

III. GBIF 소개 및 활동

GBIF(<http://www.gbif.org>)는 지구상에 존재하는 생물 개체(organisms)에 대한 1차 정보를 웹을 통하

여 제공하고 국제적인 데이터베이스 네트워크를 개발하여 세계의 생물다양성데이터가 자유롭고 널리 이용될 수 있도록 하는 것을 목적으로 OECD에서 인준하여 1999년 설립되었다.

GBIF 생물다양성데이터 네트워크는 크게 GBIF 포털, 국가거점노드(Participant Node), 데이터 노드(Data Node)로 구성되고 각 노드의 역할은 <그림 1>와 같다.



▶▶ 그림 1. GBIF 노드의 역할

한국의 경우 과학기술부의 지원으로 KBIF (Korean Biodiversity Information Facility)로 참여하고 있고 IT 기술적 인프라를 갖춘 KISTI에서 국가거점노드의 역할을 수행하고 있다[6].

GBIF 데이터 네트워크 아키텍처는 XML, SOAP(DiGIR, BioCASE), WSDL, UDDI 등의 스택을 가지는 웹 서비스(Web Services) 모델을 지향하고 있다. 웹 서비스는 분산되고 이질적인 데이터베이스를 연결할 때 XML 문서를 교환하기 때문에 노드별로 분리된 기술적용, 다국어 지원 등의 장점이 있다.

GBIF는 위에서 설명한 DarwinCore/DiGIR, ABCD/BioCASE의 표준 활동과 소프트웨어 패키지 개발에 참여하고 있으며 데이터 노드에 관련 소프트웨어 패키지를 웹을 통하여 보급하고 데이터 제공자 소프트웨어 설치, 사용방법 등에 대한 기술교육 워크숍 등을 실시하고 있다. 이와 같은 노력으로, 전 세계

생물다양성데이터를 검색할 수 있는 데이터 포털 (<http://www.gbif.net>)이 2004년 3월에 개설되었고 2005년 4월 현재 109개 데이터 제공자의 약 6,700만 건의 데이터가 웹으로 서비스되고 있다.

IV. 한국생물다양성데이터 검색포털 구축

현재 국내에는 여러 기관에서 다양한 종류의 생물다양성 데이터베이스가 구축되어 서비스되고 있고 앞으로 박물관, 과학 관 등의 데이터가 디지털화 될 것으로 예상된다.

KISTI에서는 2002년 19개의 생물다양성 데이터베이스(<http://biodiversity.kisti.re.kr>)를 구축하여 웹을 통하여 서비스하고 있고 한국생명공학연구원 생물자원센터, 국립수목원, 국립중앙과학관, 대학 박물관 등 많은 기관에서 특색에 맞는 서비스를 하고 있다. 그렇지만 현재 국내에 생물다양성데이터 교환 표준과 프로토콜이 없어 국내 생물다양성 데이터베이스를 네트워크로 연결하는 것은 쉽지 않은 일이다. 따라서 국내의 생물다양성 관련 전문가들이 모여 데이터를 교환, 활용할 수 있는 방안을 모색해야 할 것으로 사료된다.

1. 국내 데이터노드 구축

국내 생물다양성데이터의 데이터표준 및 교환 프로토콜이 없기 때문에 현재 몇몇 연구 기관(KISTI, KRIBB, 국립중앙과학관)을 중심으로 GBIF의 생물다양성 데이터 네트워크 구축에 참여하면서 DarwinCore데이터 표준과 DiGIR 프로토콜을 국내 데이터에 적용하여 데이터를 공유, 서비스하고 있다.

GBIF의 한국의 국가거점노드 역할을 수행하고 있는 KISTI는 국내생물다양성 네트워크를 구축하기 위하여 KBIF(<http://www.kbif.re.kr>) 웹사이트를 구축 운영하면서 국내의 잠재 생물다양성데이터노드(정부 유관기관, 대학 및 사설 박물관 등)에게 국제적인 생물다양성기구 및 GBIF의 동향을 알리고 데이터

노드 구축에 필요한 표준 데이터 형식, 소프트웨어 도구(DarwinCore/DiGIR Package, ABCD/BioCASE, Data Repository Tools)의 보급과 관련 교육을 실시하고 있다.

2. 생물다양성데이터 검색포탈 구축

분산된 자원을 통합 검색 또는 활용될 수 있도록 많은 프로토콜(RPC, CORBA, DCOM, Web Services)과 아키텍처가 개발되어 왔다. 그렇지만 이러한 활동이 어느 특정 분야와 결합하기 위해서는 그 분야의 내용을 충실히 반영할 수 있어야 한다. 즉, 생물다양성분야에서는 DarwinCore, ABCD스키마 형식의 데이터를 지원해야 한다.

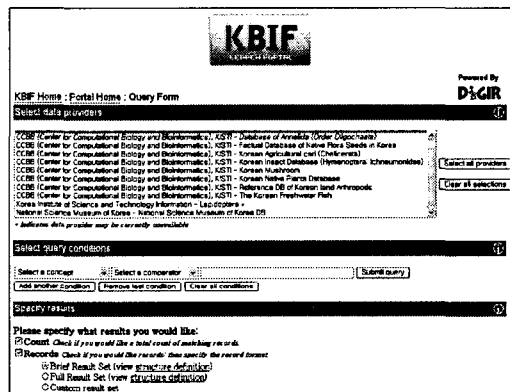
DarwinCore는 DiGIR 프로토콜을 사용하면 되는데 SourceForge.net(<http://sourceforge.net/projects/digir>)에는 현재 DiGIR 프로토콜을 사용하는 자원을 검색할 수 있도록 DiGIR노드와 DiGIR 검색포탈 소프트웨어를 오픈소스의 형태로 개발하고 있다. 검색포탈 소프트웨어는 포탈엔진(Portal Engine)과 포탈표현(Portal Presentation) 계층으로 크게 분리되어 있고 MS 윈도우 환경에서 개발되고 있다.

KISTI에서 이번에 구축한 국내생물다양성데이터 검색포탈[7]은 이 소프트웨어를 기반으로 하여 몇몇 내용을 필요에 맞게 수정하여 Linux 운영체제에 운영될 수 있도록 하였고 현재 위에서 설명한 3개 기관의 10개 자원을 대상으로 데이터 검색을 제공하고 있고 지속적인 기술보급을 통하여 향후 다른 기관의 데이터를 검색할 수 있게 할 예정이다. <그림 2>는 검색포탈화면이다. 검색포탈은 데이터노드가 추가 될 때마다 소스코드의 수정 없이 설정파일로 데이터노드를 조정할 수 있어 확장성이 뛰어난 것이 장점이다.

이번에 구축된 검색포탈은 기존의 개별 사이트에서 제공하는 데이터를 통합하여 검색할 수 있어 국내의 생물다양성데이터의 분포 및 현황을 한눈에 파악할 수 있을 것으로 기대되고 데이터를 직접 프로그래밍적으로 접근할 수 있기 때문에 2, 3차의 부가가치 포

로그래밍을 창출할 수 있을 것으로 기대된다. 앞으로 검색포탈 사이트의 기능 및 내용 보강, 한글화 진행, 사용자에게 친숙한 화면 제공 등을 추진할 예정이다.

한국의 거점노드인 KISTI는 2005년 4월 아시아권 역의 GBIF 데이터서비스 및 미러사이트 구축 사업에 선정되었다. 이를 계기로 KISTI는 GBIF와 기술협력 및 국내 생물다양성데이터 유통 및 서비스에 최선을 다할 계획이다.



▶▶ 그림 2. KBIF 검색포탈

V. 결론

지금까지 생물다양성데이터 네트워크 구축 및 유통을 위해 정보화의 관점에서 생물다양성데이터 표준(DarwinCore, ABCD 스키마)과 데이터 교환 프로토콜(DiGIR, BioCASE), 그리고 지구상의 생물다양성 데이터를 인터넷을 통하여 자유롭게 널리 이용될 수 있도록 관련 표준과 소프트웨어 개발 등에 활발한 활동을 하고 있는 GBIF에 대해 알아보았다. 또한 GBIF의 활동에 참여하면서 국내의 생물다양성데이터를 활용할 수 있도록 국내 데이터노드 구축을 힘쓰는 거점노드(KISTI)와 검색포탈 구축에 대해서 서술하였다.

전 세계의 생물다양성데이터를 공유하고 활용하려는 움직임이 GBIF를 중심으로, 국내에서는GBIF의

한국 거점 노드인 KISTI 그리고 한국생명공학연구원의 생물자원정보센터를 중심으로 지금 활발하게 시작되고 있다[8]. 국내에서도 생물다양성데이터를 보유하고 있는 정부의 유관기관, 대학 및 사설 박물관, 개인 소장자 등 관련자들이 데이터를 널리 활용하려는 자구적 프로젝트에 세심한 관심을 가지고, 국내의 생물다양성 연구 분야의 지속적 발전을 통한 국가적 차원의 생물자원 주권 확립에 적극적으로 동참하는 노력이 필요한 때이다.

현재 선진 외국과 국내의 생물다양성데이터의 활용, 유통의 기술 격차는 5년 정도 격차가 있을 것으로 생각된다. 국내의 생물다양성데이터 인프라와 국내 데이터노드가 구축되고 생물다양성데이터의 공유 및 활용의 인식이 확산되면 조만간 국제 사회에서 한국의 위상을 높이고 관련 분야를 선도해 나갈 수 있으리라 사료된다.

■ 참고 문헌 ■

- [1] James L. Edwards, "Research and Societal Benefits of the Global Biodiversity Information Facility", *BioScience*, Vol. 54 No. 6, June 2004
- [2] DarwinCore, <http://speciesanalyst.net/docs/dwc/index.html>
- [3] ABCD, <http://bgbm3.bgbm.fu-berlin.de/TDWG/CODATA/Schema/default.htm>
- [4] DiGIR, <http://digir.net>
- [5] Markus Doring, Javier del Torre, "BioCASE, Provider Software User Guide", Botanic Garden and Botanical Museum, Berlin Dahlem, <http://www.biocase.org/dev/provider/>
- [6] KBIF, <http://www.kbif.re.kr>
- [7] KBIF Search Portal, <http://dataprotider.kbif.re.kr:10080/pres/PresentationServlet?action=home>
- [8] 박형선, 안성수, 권창혁, 양진호, "생물다양성데이터 활용을 위한 국가 데이터노드(KBIF) 구축", 지식정보인프라, 통권 16호, 2004년 10월