XML기반 생물자원정보 관리시스템 설계

양진호[0] 이계준 안부영 박형선
한국과학기술정보연구원
{spearjin, kjlee, ahnyoung, seonpark}@kisti.re.kr

Design of Biological Resource Information Management System based on XML

Jin-Ho Yang[0], Kye-Jun Lee, Bu-Young Ahn, Hyung-Seon Park
Korea Institute of Science and Technology Information

요 약

생물다양성정보는 크게 종(species) 정보와 내용(content) 정보로 나눠 데이터베이스화한다. 이때, 분야별 데이터베이스 통합은 종에 대한 횡적인 통합과 내용에 대한 종적인 통합이 동시에 이루어지며, 분류학자들에 의해 정의된 항목과 국제적인 GSD(Global Species Database) 구축의 표준이 되는 내용들을 기반으로 구성요소들의 표준화 정의와 XML기반 표준 DTD를 작성한다. 또한 분산되어 있는 데이터베이스를 대상으로 통합 검색이 가능하도록 Mediator기법을 적용하여 관리시스템을 설계하였다.

1 서 론

국내에 지금까지 구축된 생물다양성 정보들이 제공하는 데이터들은 질의 형식, 데이터 모델, 스키마 구조, 사용하는 시스템에서 이질적인 특성들이 나타난다. 이는 정보 자원들의 통합을 어렵게 하는 분산성(distribution), 자치성(autonomy), 이질성(heterogeneity), 모호성(ambiguity)의 요인으로 작용한다. 그리고, 분산환경 하에서 데이터의 분산된 형태를 보면 횡적분산(horizontal distribution)과 종적분산(vertical distribution)의 형태가 있는데 현재까지 구축된 정보를 통합할 때에 이 두 가지 모두를 해결해야 하는 어려움을 가지고 있다.

이와 같은 문제점을 극복하기 위해서는 생물자원 정보의 교환 및 공유가 가능하도록 하기 위한 표준화 작업이 선행되어야 한다. 생물자원정보의 표준화를 통한 원시 데이터 구축으로 국가내 생물정보의 정형화 및 상호연계를 통한 확장성 제공과 가공을 통한 정보의 재사용과 서비스의 질적인 향상을 기초로 하여 다양한 연구분야에 적용하므로 연구활성화 및 성과의 극대화가 가능해 진다. 본 논문에서는 이와 같은 표준의 기반으로 W3C(World-Wide Web Consortium)의 인터넷 전자문서 표준인 XML(eXtensible Markup Language)을 사용한다.

한편, 각기 구축하여 왔던 생물자원 데이터베이스를 궁극적으로 공유하기 위해서는 이들 데이터베이스를 통합 할 필요성이 있다. 그러나, 지금까지 구축되어 왔던 데이터베이스를 물리적인 하나로 통합하는 것은 기존에 투자되었던 노력과 예산을 낭비할 뿐만 아니라 각 부처 또는 기관이 해당 분야의 생물자원의 정보를 수집하고 유지해 오던 전문성을 살리지 못하게 되는 단점이 있게 된다. 따라서, 본 논문에서는 각기 구축된 데이터베이스의 독립성을 최대한 유지하면서 사용자는 이질적(heterogeneous)이며, 분산(distributed)되어 있는 데이터베이스에 대해 투명(transparent)하게 마치 하나의 통합된 데이터베이스처럼 사용할 수 있는 Mediator 방식의 통합 방식을 기반으로 한다[1][2].

생물다양성정보는 크게 종(species) 정보와 내용(contents) 정보로 나눠 데이터베이스화한다. 이때, 분야별 데이터베이스 통합은 종 정보에 대한 횡적인 통합과 내용 정보에 대한 종적인 통합이 동시에 이루어지며, 분류학자들에 의해 정의된 항목과 국제적인 GSD(Global Species Database)구축의 표준이 되는 내용들을 기반으로 구성요소들의 표준화 정의와 XML기반 표준 DTD를 작성한다. 지역(local) 표준 데이터베이스 구축을 위해 표준화가 정의된 DTD를 기반 입력시스템에 정보를 입력한다. 입력시스템은 컴포넌트(component) 형식으로 만들며, 자동으로 입력시스템을 생성하게 된다. 표준화를 기반으로 통합 DB를 구축하고 사용자들에게는 분산통합검색해서 Real-Time으로 결과를 제공한다. 또한 정보 제공자들에 의한 검증까지 전체적인 기능을 갖는 관리시스템을 설계하였다.

2 관련연구

각 분야의 생물자원정보에 대한 DB를 구축하고 독립적으로 서비스하던 기존의 방법에서 벗어나 범 세계적으로 존재하는 생물다양성정보를 국가차원으로 네트워크화 하여 방대한 양의 정보를 관리·공유·검색하고 이를 이용함으로서 경제·환경·사회적 편익을 증대하고자 하는 목적에 초점을 맞추어 가는 것이 대부분이다.

2.1 국외연구

2.1.1 GBIF(Global Information Facility)

GBIF는 국제생물다양성 상호협력 Project로서 다자간의 (비)투표회원국가와 국제기구들간의 상호동의를 통해 MoU에 서명함으로서 회원국가별로 자국내 대표중심점(National Node)을 구성하고 자국의 정보를 공유할 수 있는 환경을 제공하도록 유도하며, 각국에 존재하는 다양한 형태의 생물다양성 데이터베이스를 네트워크화 하여 상호 이용함으로서 경제·환경·사회적 편익증대를 도모하기 위한 국제 생물다양성정보기구이다.[3].

2.1.2 Species 2000

Species 2000은 지구상의 모든 알려진 종들의 데이터를 포함하는 GSD(Global Species Database) 구축을 목적으로 하고 있으며, 지구상의 식물, 동물, 균류, 미생물에 대한 과학적인 이름, 상태, 분류 등의 정보를 포함하는 데이터베이스 종 체크리스트를 통해서 사용자들이 종의 규명이 가능한 색인을 구축하고 있다.

'Catalogue of Life'는 Common Name, Scientific Name, Reference에 의한 검색을 Annual CheckList와 Dynamic CheckList로 구분하여 검색 서비스를 제공하고 있다. 이 시스템에서 제공하는 기능 중에서 다른 시스템들과 구분되는 것은 Update, Add, Get XML File 등의 기능으로 검색 결과에 대한 수정, 새로운 정보에 대한 추가와 XML 표준 문서를 제공으로 데이터베이스의 양적인 확장뿐만 아니라 깊이 있는 정보의 구축을 가능하게 한다는 점이다. 또한, 정보의 제공자와 관련 참고자료등의 정보와 검색 결과의 자세한 정보를 보기 위해 다른

독립적인 생물다양성 정보를 구축해놓은 유관 데이터베이스들과의 연계를 통한 통합 검색의 가능을 제공하는 것이다[4].

2.1.3 ITIS(Integrate Taxonomic Information System)

ITIS는 2001년 6월에 ITIS(Integrated Taxonomic Information System)과 'Catalogue of Life'의 생성을 위해 참여하고 있다. 동물, 식물, 균류, 미생물 분야를 대상으로 Taxo 정보를 데이터베이스화하고 현재 서비스를 수행하고 있으며, Taxo 정보뿐만 아니라 관련정보까지 입력이 가능한 입력 시스템을 구축하여 운영하고 있다[5].

2.2 국내연구

현재 국내의 생물다양성 정보들은 국내의 고유종에 대한 정보에서부터 자생종에 이르기까지의 정보를 대상으로 각 기관에 종속적으로 독립적인 구축, 관리, 서비스가 이루어지는 것이 대부분이다[6]. 그러므로 각 DB들은 여러 기관에 중복되어 있으며, 국내 데이터베이스들은 대부분 적은 규모로 서비스를 하고 있고 정보의 공유나 연계가 거의 이루어지지 못하기 때문에 하나의 완전한 정보를 얻기 위해서는 여러곳의 DB를 검색하여 사용자들이 조합하는 과정을 거쳐야만 원하는 정보 전체를 얻을 수 있다. 또한, 데이터를 보유하고는 있지만 정보화에 대한 지식과 인력의 부족으로 DB화하지 못하여 서비스되지 못하는 정보들이 많은 실정이다.

생물다양성정보는 멀티미디어 정보로서 텍스트, 이미지, 동영상, GIS정보들로 구성되어 있으나 현재 데이터베이스 대부분은 이러한 정보를 고루 갖추고 있지만 기술적인 적용이 미비한 실정이다. 국내에는 대학, 연구소, 협회, 기업등을 기반으로 생물다양성 정보에 대한 내용 검색, 사이버 박물관, 자연사 박물관, 종자은행 등의 다양한 서비스가 이루어지고 있으나 규모나 내용 면에서 내실을 다져야 할 필요성을 가지고 있다.

국가적인 경쟁력을 가지기 위해서는 유관기관들간의 네트워크를 통한 정보 공유와 표준화를 기반으로 하는 통합 추진으로 Macro와 Micro의 두 가지 모두를 겸비한 DB구축이 필요하다. 이를 바탕으로 세계적인 네트워크 체계를 구축하므로 생물자원 부국(magadiversity)의 우위를 가질 수 있게 될 것이다.

3 관리시스템 설계

3.1 생물자원정보 DTD 설계

생물자원정보의 표준화를 위해 각 구성요소를 대상으로 DTD를 설계하였다. DTD는 종 정보와 내용 정보로 나누어 작성하였으며, 종 정보 DTD는 횡적인 통합을 위한 것이며, 내용정보 DTD는 종적인 통합을 고려하였다.

3.1.1 생물 종에 대한 DTD 설계

```
<?xml version="1.0" encoding="euc-kr"?>
<!-- 종(Species)정보 -->
<!-- ENTITY 선언 -->
<!ENTITY % name "(#PCDATA)">

<!-- Species에 대한 ELEMENT 선언 -->
<!ELEMENT kingdom (subkingdom | phylum | division)*>
<!ELEMENT kingdom %name; >
<!ELEMENT subkingdom (phylum | division)*>
<!ELEMENT subkingdom %name; >
    . . . . . .
    . . . . . .
<!ELEMENT family (subfamily | tribe | genus)*>
<!ELEMENT family %name; >
```

표 1 생물종에 대한 DTD

종 정보는 계·문·강·목·과·속·종에 해당하는 것으로 각각의 구분은 super, infra, sub등 세부적으로 나눠진다. 국내 생물자원정보를 국제적인 데이터베이스 구축을 위해 본 논문의 DTD는 국제적인 분류정보 검색을 제공하고 있는 Species2000과 ITIS에서 제공하는 종 정보 검색 내용을 기반으로 했으며, 국내 모든 생물종에 대한 카탈로그 구축을 목적으로 한다. 표 1은 생물 종에 대한 DTD를 작성한 것이다.

3.1.2 생물 내용에 대한 DTD 설계

생물자원정보의 일반적인 특징을 살펴보면 다음과 같다.

- 자료의 양과 범위에 대한 변화의 정보가 매우 높고 복잡하다.
- 진화적인 데이터베이스로 데이터베이스 스키마의 변화가 빠르다.
- 같은 정보에 대한 데이터 표현 방법이 상이하다.
- DB 제공자들의 전산화를 위한 데이터베이스 구조 및 스키마 구조에 대한 이해가 부족하다.
- 기본 자료뿐 아니라 의미상의 정보도 함께 제공되어야 한다.
- 자료 서비스 시에 과거의 자료에서 최신 자료까지 모두를 제공해야 한다.

위와 같은 일반적인 특징은 정보의 확장가능성, 논리적·구조적 정보까지 모든 정보에 대한 관리시스템을 요구하게 된다. 따라서 생물자원을 표현하고 설명하기 위한 내용정보는 확장성을 고려하여 DTD를 설계해야 한다.

표 2는 종 설명정보, 멀티미디어정보, 서식지정보, 참고문헌정보, 명명자정보, 관련정보, 파생정보 등으로 구성되어 진다. 종 설명정보는 생물종의 특징정보(번식방법, 수명, 용도, 생물학적특징, 구성성분), 설명정보(어원, 종설명, 특이사항), 기타정보 등을 표현한다. 이와 같이 생물자원정보 표준 DTD는 부분별 정보로 나누어져 구성이 되며, 정보의 확장 가능성이 있는 엘리먼트 타입을 ANY로 설정하여 충분한 확장성을 고려하였다.

```
<?xml version="1.0" encoding="euc-kr"?>

<!-- 한국과학기술정보연구원 생물자원정보실 -->
<!-- 본 DTD는 생물자원정보의 표준화를 위하여 작성한
것으로 컨텐트(Content)정보로 표현하기 위한 것이다.
-->

<!-- ENTITY 선언 -->
<!ENTITY % text "(#PCDATA)">

<!-- Species에 대한 ELEMENT 선언 -->
<!ELEMENT species (description, multimedia,
habitation, reference, namer, relation, derivation)*>
<!ATTLIST species ancode CDATA #REQUIRED>
```

표 2 컨텐트 전체 구성 정보

3.2 생물자원정보 입력시스템 설계

3.2.1 데이터 흐름도

생물자원정보의 입력을 위해서 우선 데이터 입력자가 입력시스템 생성기를 통해 입력대상이 되는 컴포넌트를 선택하여 입력시스템을 생성한다. 그리고, 입력은 생성된 입력시스템을 통해서만 이루어지며, 분야별 표준 DTD를 기반으로 해서 XML 문서를 생성한다. 작성된 XML문서는 XML문서 저장시스템에 의해 자동으로 분석되어 논리정보, 구조정보, 내용정보 모두를 포함하여 데이터베이스가 구축된다.

3.2.2 컴포넌트 구성

생물자원정보를 7개의 대분류와 세부적인 중분류, 소분류로 구분하였으며, 소분류는 각각의 실제 값을 가지게 되는 엘리먼트에 해당한다. 따라서, 컴포넌트는 XML의 엘리먼트에 해당하는 것이며, 분류정보는 XML의 구조정보로 표현될 수 있도록 하였다.

3.2.3 데이터 입력 방법

입력시스템을 통한 데이터 입력 방법은 다음과 같다. 첫째, 입력자는 종 정보를 종 정보입력시스템을 통해 입력한다. 입력된 정보는 중복체크를 거치며, XML문서로 작성된 후에 DB에 저장된다. 둘째, 종 정보는 DB에 저장되면서 제어번호가 부여되며 고유키를 가지게 된다. 셋째, 입력자는 종 정보입력 후에 컨텐트 정보 입력시스템에서 종에 의한 정보를 검색하게 되며,

검색결과에 해당하는 정보를 입력하게 된다. 넷째, 종 정보와 내용 정보를 기 구축된 DB를 검색하여 정보의 존재여부를 확인한 후에 정보가 있을 경우 로딩하는 기능을 포함한다. 다섯째, 입력자는 필요시에 정보를 추가할 수 있으며, 추가되는 정보는 표준 스키마의 확장에 해당하며 표준 스키마를 기반으로 각 분야별 스키마가 생성되게 된다. 여섯째, 모든 XML 문서는 하나의 통합된 DB에 구축이 되며, 이를 서비스하게 된다.

3.3 생물자원 저장시스템 설계

3.3.1 로컬정보 생성

각 분야별 생물자원정보를 구축하는 로컬에 XML문서를 생성하기까지의 과정을 보면 다음과 같다.

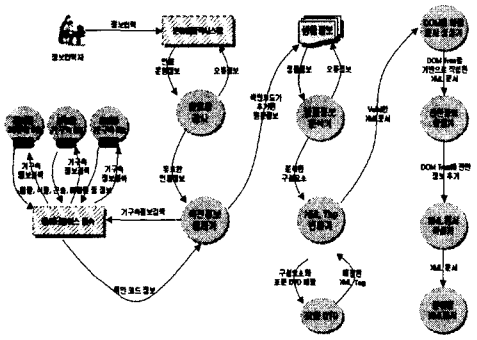

그림 1 로컬정보생성 과정

3.3.2 Mediator 기반 통합시스템

통합 DB는 물리적인 통합이 아닌 논리적인 통합으로 모든 데이터는 분야별로 로컬에 존재하여 서비스 요청이 있을 경우에만 로컬의 정보를 검색하고 서비스하는 논리적인 통합을 Mediator 기법을 적용하여 다음과 같이 설계하였다.

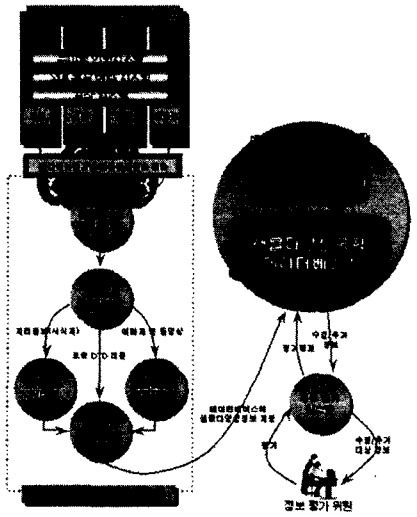

그림 2 Mediator 기반 통합시스템

4 생물자원정보 관리시스템 설계

생물자원정보는 최종적으로 다음과 같은 기능을 가져야 한다. 첫째, 실물과 정보가 하나의 연계된 시스템내에 공존해야 한다. 둘째, 국제적인 표준화를 기반으로 하는 GSD System을 구축한다. 셋째, 구축된 정보에 대한 지식화 작업 및 정보 분석 기법을 적용한다. 넷째, 여러 분야의 기술이 집적된 시스템을 설계하고 구축한다. 다섯째, 데이터 확장과 진화적인 스키마를 포함해야 한다. 이러한 관리 체계를 만들기 위한 시스템 설계는 다음과 같다.

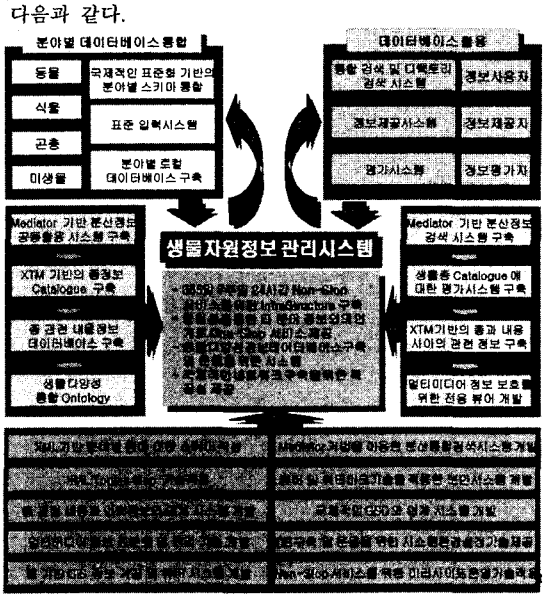

그림 3 생물자원정보 관리시스템

5 결론

생물자원정보를 효율적으로 관리하기 위한 시스템은 필요하며 표준화된 정보를 기반으로 원활한 서비스를 사용자들은 요구하고 있다. 이러한 것들을 모두 포함하고 발전시키기 위해 본 논문에서는 표준 DTD를 작성하고 이를 이용하여 컴포넌트 기반의 입력 및 저장 시스템을 설계하였다.

입력된 데이터는 XML 문서로 작성되어 구조정보, 논리정보, 내용정보 등으로 구분되어 DB화되며, 통합검색을 위해 랩퍼(wrapper)기반의 Mediator 기술을 적용하여 분산된 정보를 논리적으로 통합하여 검색할 수 있도록 하였다. 이에 많은 사람들이 참여하는 국제적인 연계가 가능한 생물자원정보 관리시스템을 설계하고 기능 정의를 하였다.

본 논문은 현재 생물자원정보화를 위해 필요시 되어 왔던 표준화, 데이터베이스 구축, 정보관리, 정보유통을 위해 반드시 요구되는 사항들에 대해 설계하였으며, 설계된 내용을 기반으로 시스템을 구축하여 현재까지 문제점으로 지적되어오던 많은 부분을 해결하고자 한다.

참고 문헌

[1] 류근호, "생물다양성 정보의 투명한 접근을 위한 데이터베이스 분산 통합 방안," 한림심포지움, 2002
[2] 이규철, "XML 기반 Topic Map과 Mediator를 이용한 생물다양성 정보 통합과 체계적 분류 및 검색," 한림심포지움, 2002
[3] GBIF(The Global Biodiversity Information Facility), http://www.gbif.org/index.html
[4] Species 2000, http://www.sp2000.org
[5] Integrated Taxonomic Information System, http://www.itis.usda.gov
[6] KISTI 생물자원정보 네트워크, http://biodiversity.kisti.re.kr