# DADI 기반의 생물다양성정보에 대한 GRM 구축

## (Contracture for GRM of Biological Resources Information of based DADI)

이 계 준* 박 형 선** 안 부 영** 양 진 호**
(KyeJun Lee) (Hyung-Seon Park) (Bu-Young Ahn) (Jin-Ho Yang)

**요 약** 본 논문에서는 첫째, 생물자원정보 데이터베이스는 크게 생물종 정보 구축과 종정보를 대상으로 구축되어지는 컨텐트(content) 정보로 나눠 XML(eXtensible Markup Language)을 기반으로 데이터베이스화하는 것이다. 둘째, 분류학자들에 의해 정의된 항목과 국제적인 GSD(Global Species Database) 구축의 메타데이터가 되는 항목들을 중심으로 정보가 구축되어야 하며, 효율적인 지역(Local) 정보의 데이터베이스화를 위하여 컴포넌트(Component) 기반의 입력시스템을 구축하여 제공. 셋째, 정보의 서비스 및 공동활용 체제를 구축하기 위하여 DADI(Data Access and Data Interoperability) 기반의 GRM(Global Road Map)을 구축의 3단계 과정을 통해 생물자원정보에 대한 데이터베이스를 구축하고 원활한 서비스 체제 구축을 위한 연구를 수행하였다.

**Abstract** In this paper consisted of three steps for the research, The first, The Database of Biological Resource Information are constructing for species Information and Content Information of based XML. The second, The item of defined from the analysts and must be considered for national GSD(Global Species Database), Supply and Contracture of Input System of based Component for the Efficient Local Information Database. The third, Information Service and Interoperability are using the GRM(Global Road Map) of based DADI. These are able to accomplish to Contracture for Database and Service structure of Biological Resources Information.

## 1. 서 론

정보는 생성에서 소멸까지의 과정에서 가공 및 재구성을 통해 가치 있는 의미를 부여되며, 다양하고 효과적인 서비스 제공을 위한 연구가 많은 부분에서 이루어지고 있다. 시대의 흐름과 환경의 변화가 연구 동향이나 가치관에 영향을 미치고 있기 때문에 국내의 생물다양성 혹은, 생물다양성에 대한 분야의 인식이 높아져 많은 투자와 연구가 각 관련 분야에서 활발해 졌다.

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 생물다양성 정보의 교환 및 공유가 가능하도록 하기 위한 표준화 작업이 선행되어야 한다. 정보의 표준화를 통한 원시 데이터 구축으로 국가 내 생물정보의 정형화 및 상호연계를 통한 확장성 제공과 가공을 통한 정보의 재사용과 서비스의 질적인 향상을 기초로 하여 다양한 연구 분야에 적용하므로 연구 활성화 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 사용을 제

* 한국과학기술정보연구원 생물자원정보실 연구원
** 한국과학기술정보연구원 생물자원정보실 선임연구원

안한다.

생물다양성정보는 크게 생물종 정보 구축과 종정보를 대상으로 구축되어지는 컨텐트정보로 나눠 데이터베이스화한다. 이때, 분야별 데이터베이스 통합은 종정보에 대한 횡적인 통합과 컨텐트에 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 만들어 진다. 본 논문에서는 생물다양성정보 DB 구축에 꼭 맞는 컴포넌트 기반 입력시스템을 구축을 제안한다.

각기 구축하여 왔던 생물다양성 데이터베이스를 궁극적으로 공유하기 위해서는 이들 데이터베이스를 통합하여야 할 필요성이 있다. 그러나 지금까지 구축되어 왔던 데이터베이스를 물리적으로 하나의 데이터베이스로 통합하는 것은 기존에 투자되었던 노력과 예산을 낭비할 뿐 만 아니라 각 부처 또는 기관이 해당 분야의 생물다양성의 정보를 수집하고 유지해 오던 전문성을 살리지 못하게 되는 단점이 있게 된다. 따라서 본 논문에서는 각기 기구축된 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 통합 방식을 제안한다.

마지막으로 본 논문에서는 생물다양성정보에 대한 접근 및 상호운용을 위하여 DADI(Data Access and Data Interoperability)를 고려한 GRM(Global Road Map) 작성을 제안한다.

## 2. 구현사례 선정 배경

### 2.1 GBIF(Global Biological Information Facility)

국가나 국제기구들 사이에 MoU를 맺고 분산되어 있는 국제적인 생물다양성정보 DB를 네트워크 상에서 공동활용 하고자함을 목적으로 함. 또한, 인터넷상에서 모든 생물다양성인 종정보와 생태정보 데이터들에 대해 쉽게 접근할 수 있도록 하기 위해 현재 추진 중에 있으며, 생물다양성과 생태정보 분야의 복잡하고 사회적으로 중요한 정보들을 동적으로 접근할 수 있도록 하기 위해 정보네트워크 구축과 최신 기술을 적용한 저작도구들을 개발하여 제공하고 있다.

### 2.2 Species 2000

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

현재 Species 2000은 GBIF내에 참여하여 관계를 가지고 있으며, 'Catalogue of Life'의 생성을 위해 참여하고 있다.

'Catalogue of Life'는 Common Name, Scientific Name, References에 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 점이며, 정보의 제공자와 관련 참고자료 등의 정보와 검색 결과의 자세한 정보를 보기 위해 다른 독립적인 생물다양성 정보를 구축해 놓은 유관 데이터베이스들과의 연계를 통한 통합 검색의 가능을 제공하는 것이다.

### 2.3 ITIS(Integrated 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 Taxo정보를 데이터베이스화하고 현재 서비스를 수행하고 있으며, Taxo 정보뿐만 아니라 관련정보까지 입력이 가능한 입력시스템을 구축하여 운영하고 있다. 또한, 구축된 Taxo정보는 Taxo정보, Common정보를 기반으로 서비스를 다음과 같이 구조적인 트리 형태로 수행하고 있다.

## 3. 연구 내용

### 3.1 생물자원정보 DTD 설계

생물자원정보의 표준화를 위해 각 구성요소를 대상으로 DTD를 설계하였다. DTD는 종정보(Species)와 내용정보(Content)로 나누어 작성하였으며, 종정보 DTD는 횡적인 통합을 위한 것이며, 내용정보 DTD는 종적인 통합을 고

려하였다.

### 가. 생물종에 대한 DTD 설계

종정보는 계·문·강·목·과·속·종에 해당하는 것으로 각각의 구분은 super, infra, sub 등으로 세부적으로 나눠진다. 국내 생물자원정보를 국제적인 데이터베이스 구축을 위해 본 논문의 DTD는 국제적인 분류정보 검색을 제공하고 있는 Species2000과 ITIS에서 제공하는 종정보 검색 내용을 기반으로 했으며, 최소한 7개의 구분은 꼭 정보를 가지도록 설계하였으며, 국내 모든 생물종에 대한 카탈로그(Catalogue)를 구축하는 것을 목적으로 한다.

### 나. 컨텐트에 대한 DTD 설계

생물자원정보의 일반적인 특징은 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.
- 복잡한 자료의 검색이 가능한 질의 정의가 요망된다.
- 자료 서비스 시에 과거의 자료에서부터 최신 자료까지 모두를 제공해야한다.

위와 같은 일반적인 특징은 정보의 확장가능성, 논리적·구조적 정보까지 데이터베이스 구축, 구축되는 정보에 대한 관리시스템을 요구하게 된다.

따라서 생물자원을 표현하고 설명하기 위한 컨텐트정보는 확장성을 고려하여 DTD를 설계한 것은 다음과 같다.

첫째, 컨텐트정보는 종설명정보, 멀티미디어정보, 서식지정보, 참고문헌정보, 명명자정보, 관련정보, 파생정보로 구성되어 진다.

둘째, 종설명정보는 생물종의 특징정보(번식방법, 수명, 용도, 생물학적특징, 구성성분), 설명정보(어원, 종설명, 특이사항), 기타정보 등을 표현한다.

셋째, 멀티미디어정보는 이미지(그림, 사진)나 동영상(소리, 동영상) 등의 정보를 표현하기 위한 것으로 각각의 파일명, 타입, 파일크기의 정보를 가진다.

넷째, 서식지정보는 생물이 서식하는 서식지의 지리적인 위치 정보를 가지며 크게 위경도와 주소정보로 구분하였다. 서식지 정보는 GIS 시스템을 구축하고 서비스하는데 중요하기 때문에 정확한 정보 구축이 필요하다.

다섯째, 참고문헌정보는 종정보를 동정하기 위해 참고한 논문, 보감, 서적 등의 정보를 포함한다.

여섯째, 명명자정보는 종정보를 명명한 사람의 이름, 연락처, 전공, 국가 등의 정보를 포함한다.

일곱째, 관련정보는 종정보와 관련된 데이터베이스나 홈페이지 정보 등에 해당하는 것으로 상세 정보, 유사 정보, 관련 정보 등을 포함된다.

여덟째, 파생정보에 대한 관련 정보(기생식물의 기주가 되는 식물 : host, 기주식물의 파생식물 : associate)등이 포함된다.

## 3.2 생물자원정보 입력시스템 설계

생물자원정보의 정형화된 데이터 구축을 위하여 작성된 DTD의 엘리먼트를 기반으로 입력시스템을 설계하였다.

### 가. 데이터 흐름도

생물자원정보의 입력을 위해서 우선 데이터입력자가 입력시스템 생성기를 통해 입력대상이 되는 컴포넌트를 선택하여 입력시스템을 생성한다. 그리고, 입력은 생성된 입력시스템을 통해서만 이루어지며, 분야별 표준 DTD를 기반으로 해서 XML문서를 생성한다. 작성된 XML문서는 XML문서 저장시스템에 의해 자동으로 분석되어 논리정보, 구조정보, 내용정보 모두를 포함하여 데이터베이스가 구축이 되는 과정은 다음과 같다.

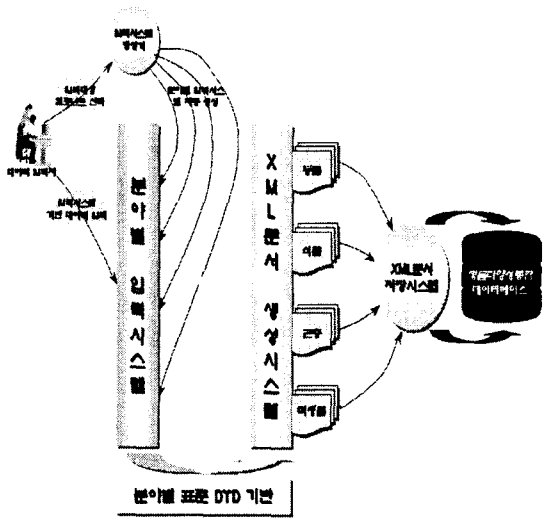


그림 1 입력시스템의 데이터 흐름도

### 나. 컴포넌트 구성

생물자원정보를 7개의 대분류와 보다 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.

다음은 컴포넌트의 전체 구성이다.

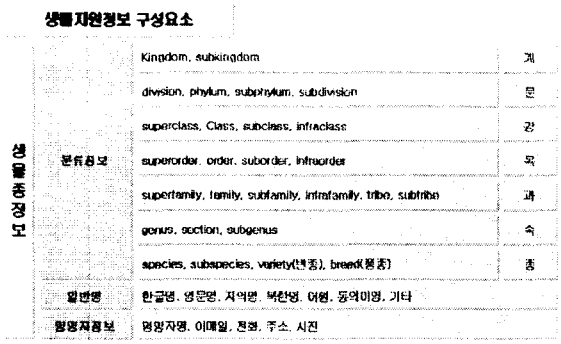


그림 2 생물종 컴포넌트

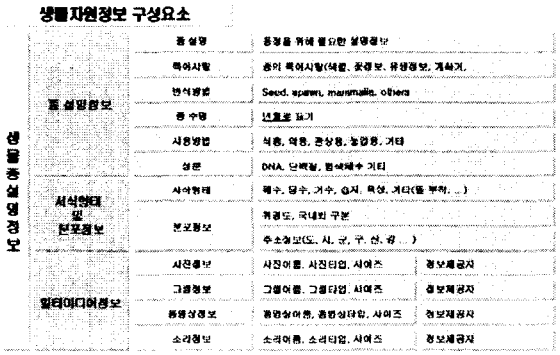


그림 3 생물종 설명 컴포넌트

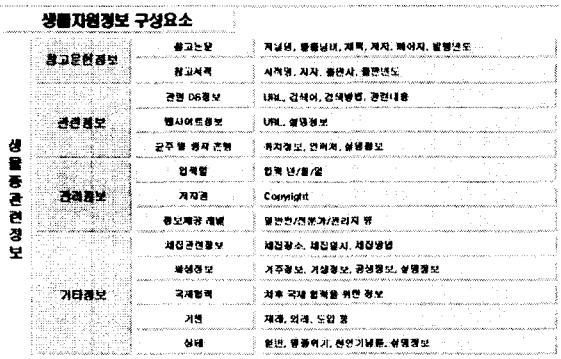


그림 4 생물종 관련 컴포넌트

### 다. 데이터 입력 방법

입력시스템을 통한 데이터 입력 방법은 다음과 같다.

첫째, 입력자는 종(Species)정보를 종정보입력 시스템을 통해 입력한다. 입력된 정보는 중복 체크를 거치며, XML문서로 작성된 후에 데이터베이스에 저장된다.

둘째, 종정보가 데이터베이스에 저장이 되면서 제어번호(ancode)가 부여되며 고유키(Primary Key)를 가지게 된다.

셋째, 입력자는 종정보입력 후에 컨텐트정보를 입력해야 하는데 이때, 컨텐트정보 입력시스템에서 종에 의한 정보를 검색하게 되며, 검색 결과에 해당하는 정보를 입력하게 된다.

넷째, 종정보와 컨텐트정보는 기구축된 데이터베이스를 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 데이터베이스에 구축이 되며, 이를 서비스하게 된다.

## 3.3 생물자원 저장시스템 설계

각 분야별로 분산되어 작성된 XML문서를 통합하여 저장하기 위하여 미디에이터(Mediator)기법을 적용하여 시스템을 설계하였다.

### 가. 로컬정보 생성

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

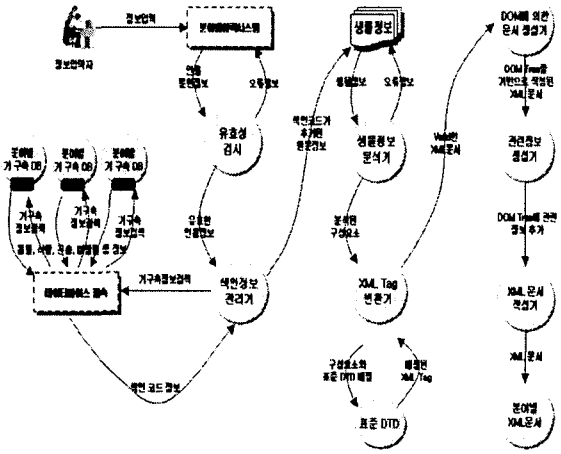


그림 5 로컬정보생성 과정

### 나. 미디에이터 기반 통합시스템

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

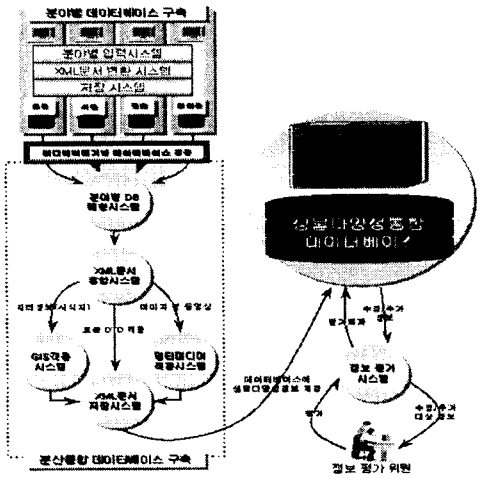

그림 6 미디에이터 기반 통합시스템

## 4. GRM 기반 네트워크

GRM은 분산되어 있는 정보에 대한 접근을 위한 정보를 가지고 있어 사용자들의 서비스 요청이 있을 경우에 Mediator 시스템을 통해 GRM의 정보를 검색하고 결과를 가지고 로컬 정보를 검색하게 된다. 또한 로컬 정보에 대한 전체적이 Local Map이 존재하고 이것을 모두 포함하는 GRM이 만들어진다.

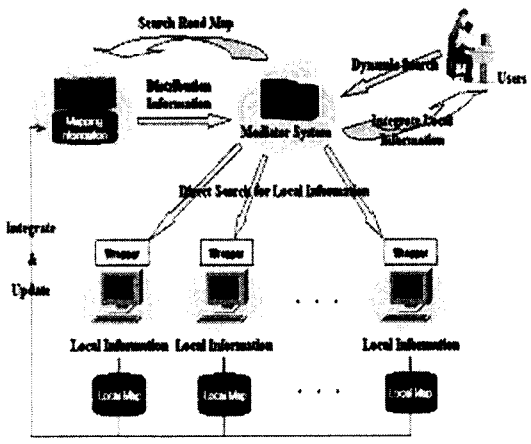


그림 7 GRM 시스템의 전체 구성

GRM 시스템은 로컬의 정보를 논리적인 관계(Association)에서 얻어지는 계층구조를 기반으로 만들어지며, 정보들 간의 관계를 정의하기 위한 기본 개념은 다음과 같다.

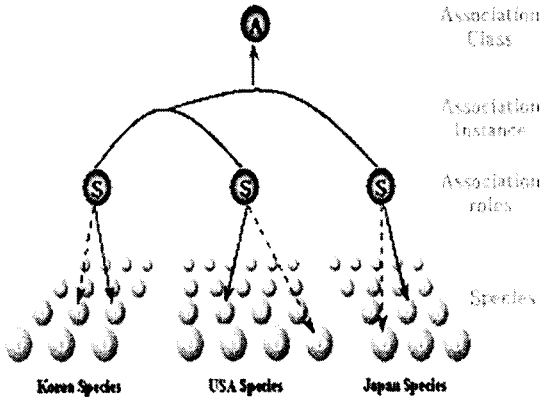


그림 8 정보들 간의 관계 정의를 위한 기본 개념

GRM은 각각의 개별적인 개체(Object)들로부터 시작을 하며, 이러한 개체들 간에 정보로서의 관련성이 같은 것들 간에 하나의 의미 부여를 통해 계층적인 구조를 가지도록 설계하는 것이다. 이것은 정보의 확장이 보장되게 되며, 원하는 정보를 다시 재정의 하여 자신의 시스템에 사용할 수 있도록 한다. 따라서 본 개념을 기반으로 해서 GRM을 생물자원정보에 적용하여 계층적으로 나타내면 다음과 같은 단계로 구분할 수 있다.

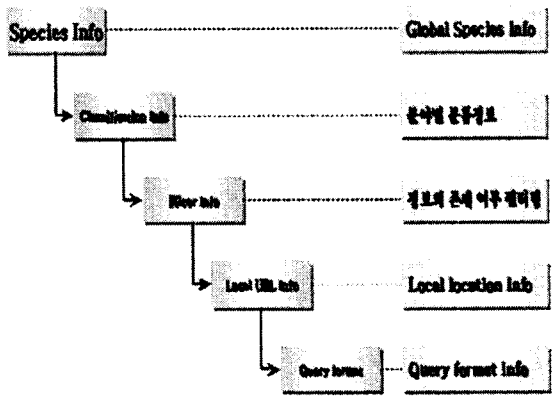


그림 9 GRM 정보 표현의 구성도

첫째, Species Info는 종정보인 종/속/과/목/강/문/계와 일반명(Common Name : 한글명, 영문명 등)을 기준으로 구성된다. 둘째, Classification Info는 종정보 이외의 정

보를 보다 세분화함으로서 검색의 효율과 분산화를 위한 분류정보이다. 셋째, Filter Info는 검색 대상이 되는 내용이 존재하는가의 정보를 가지고 있어 정보의 검색 시에 필터링 해주는 기능을 수행한다. 넷째, Local URL Info는 분산되어 존재하는 Database의 위치 정보를 가지고 있어 Global 정보와 Local 정보와의 Mapping 기능을 수행한다. 다섯째, Query Format는 Local 정보를 검색하기 위해 각각의 Database에 맞는 질의 형태로 변환하여 검색하고 결과를 반환해 준다.

## 6. 결 론

GRM 기반의 생물자원정보 네트워크를 구축하기 위해서 수행되어야 하는 부분으로 크게 나누어볼 수 있다.

첫째, 생물자원정보 데이터베이스 구축의 두 가지 목적은 첫째, 생물종자원(Species) 정보 구축, 둘째, 이것들을 기초로 하여 내용정보, 이미지정보, 동영상정보, GIS정보, 관련정보 등을 구축한다. 이것은 국제적인 흐름에 따르는 것이며, 국내에서 반드시 구축되어야 하는 Catalogue 정보와 관련 정보를 동시에 구축할 수 있는 것이다.

둘째, Catalogue정보를 구축하기 위해서는 각 분야 전문가들과 공동으로 공동활용과 검색을 위한 표준화된 구성요소를 추출하고 각 분야의 정보를 모두 포함할 수 있는 표준 DTD 작성한다. 또한 이렇게 만들어진 정보들의 따르는 로컬 시스템을 구축하는데 여기에는 입력시스템과 Global 시스템과 연동이 가능한 모듈을 포함하는 시스템이 구축되어야 한다.

셋째, 로컬시스템을 통해 각각의 분야에서 만들어지는 정보들의 표준화가 이루어지고 이렇게 데이터베이스화된 정보를 Centralize한 개념의 하나의 데이터베이스를 구축하는 것이 아닌 Mediator기법을 이용하여 분산된 정보 Real-time검색이 가능한 시스템을 구축하여 서비스하게 된다.

넷째, 가장 선행되어야 할 것은 생물종에 대한 데이터베이스이며, 이것을 기반으로 종에 관련정보들이 구축되어진다. 이렇듯, 정보화에 대한 기준이 세워지고 하나의 종에 대한 완벽한 관련 정보까지 구축되어진 다음에 종들끼리의 Association(연관)/Relation(관련) 정보가 완벽하게 적용할 때만이 올바른 데이터베이스 구축 및 운영이 가능해 진다.

다섯째, 정보의 효율적인 접근을 통한 서비스를 위하여 GRM과 같은 정보에 대한 맵을 만들어 모든 정보의 접근을 보장하므로 정보유통을 위한 기반을 조성한다.

본 논문에서는 생물자원정보를 대상으로 국내 모든 생물자원에 대한 접근을 가능하게 하며, 표준화된 데이터베이스 구축, 분산통합을 통한 정보의 고유성 및 정보 제공으로 인한 손실을 최소화하기 위한 네트워크 체제 구축 및 필요한 시스템과 방법에 대해 전반적으로 제안을 하였다. 제안된 내용을 기반으로 정보를 구축하고 유통한다면 생물자원정보 네트워크 체제가 확립 될 것이다.

## 참 고 문 헌


[1] 국립환경연구원, "생물자원주권 확보를 위한 국립생물자원보존관의 역할", 환경의날 기념 국제세미나, 2002[1]

[2] 한림원, "국내 생물자원 정보 DB 및 네트워크 운영체계 확립", 생물자원정보 콜로퀴움, 2002

[3] 충남대학교 소프트웨어연구센터, "XML 문서 저장/검색 및 분산 문서 시스템의 설계 및 구현", 연구보고서, 2001

[4] 이계준, 조현양, 최재황, 손강렬, "효율적인 KSCI 체제 구축을 위한 XML 기반 모델 설계", 한국과학기술정보연구원, 2001 가을 정보과학회 학술발표논문집, 2001

[5] 한국전자통신연구원, "이질 자료 모델의 충돌 해결 방안 연구", 중간보고서, 2001

[6] Integrated Taxonomic Information System, http://www.itis.usda.gov

[7 ]GBIF(The Global Biodiversity Information Facility) http://www.gbif.org/index.html

[8] Species 2000 http://www.sp2000.org

[9] Center for Computational Biology & Bioinformatiocs,

[10] National Center for Biotechnology Information,

[12] 손현석, 생물정보학 (Bioinformatics)해외동향 , 한국과학기술정보연구원, 지식정보인프라지