정보관리연구, vol.40, no.1 2009, pp.217-234

스키마 통합 기반 생명정보 검색시스템(BIRS) 설계에 관한 연구

A Study on Design of Schema Integration based Biological Information Retrieval System

한 건* · 이 상 호** · 안 부 영***
Keon Han · Sang-Ho Lee · Bu-Young Ahn

차 례

초 록

컴퓨터로 옮겨 놓은 생물학 실험실에서 생명과학을 연구하는 연구자가 생명정보를 확인하려면 1차적으로 생물다양성 관련 데이터베이스에서 생명체에 관한 종정보, 생태정보, 분포정보를 검색해야 한다. 그리고 그 생명체를 구성하는 유전자 서열정보와 단백질 구조정보를 Genbank, PDB 등의 유전자/단백질 데이터베이스에서 검색해야 한다. 또한 그 생명체에 관한 학술적 내용이 수록된 학술논문까지 별도로 검색해야만 그 생명체에 관한 포괄적이고도 정확한 정보를 획득하여 연구에 활용할 수 있다. 이런 일련의 과정은 연구자에게 불편함과 함께 많은 시간이 소요됨으로 인해 연구의 효율성을 저하시키는 요인이 되고 있다. 이런 불편함을 해결하기 위하여 통합검색하기 위한 여러 방법을 분석하고, 그 중 스키마 통합을 선택하였다. 또한 스키마 통합을 위하여 각각의 데이터베이스의 스키마를 분석하고 메타데이터를 추출하여 Mediated 스키마를 설계하였다. 본 논문에서 설계한 생명정보 검색시스템(BIRS, Biological Information Retrieval System)과 인터페이스를 사용하여 생명과학을 연구하는 연구자들의 연구의 효율성을 향상시킬 수 있을 것이다.

키 워 드

바이오인포매틱스, 생물다양성, 지리정보시스템, 메타데이터, 데이터베이스, 정보검색시스템

* 한국과학기술정보연구원 지식기반실 학생연구원
(Student Researcher, Dept. of Knowledge Resources, KISTI, jingkir@kisti.re.kr)
** 한국과학기술정보연구원 지식기반실 책임연구원
(Principal Researcher, Dept. of Knowledge Resources, KISTI, shlee@kisti.re.kr)
*** 한국과학기술정보연구원 차세대연구환경개발실 선임기술원
(Senior Researcher, Dept. of Cyber Environment Development, KISTI, ahnyoung@kisti.re.kr)
• 논문접수일자 : 2009년 1월 15일
• 게재확정일자 : 2009년 3월 17일

ABSTRACT

In computer-based virtual lab, a bioscience researcher who wants to obtain bio information first uses a biodiversity-related database to retrieve information on species, ecology and distribution of an organism. The researcher also needs to access gene/protein databases such as GenBank or PDB to find information on the organism's genetic sequence and protein structure. Furthermore, the researcher should search for academic papers containing the information on the organism so that his research is based on comprehensive and accurate information. This series of activities often undermines research efficiency as it takes a lot of time and causes inconvenience on the part of researchers. To solve such inconvenience, we analyzed various methods for integrated search and chosen schema integration. In addition, we analyzed each databases and extracted metadata for designing schema integration. This paper introduces a biological information retrieval system(BIRS) using schema integration and it's interface that will increase research efficiency for bioscience.

KEYWORDS

Bioinformatics, Biodiversity, Geographic Information System, Metadata, Database, Biological Information Retrieval System

1. 서론

전 세계적으로 컴퓨터 기술이 발달함에 따라 대용량 생명정보와 정보기술이 접목된 생명정보학(Bioinformatics, 바이오인포매틱스)이라는 복합 학문이 등장하여 지속적으로 발전하고 있다. 생명과학 연구자가 생명정보 관련 정보를 확인하려면 우선, 생물다양성 관련 데이터베이스에서 생명체 종정보에 관한 검색을 하고 그 생명체를 구성하는 유전자 서열정보와 단백질 구조정보를 유전자/단백질 데이터베이스에서 검색해야 한다. 또한 그 생명체에 관한 학술적 내용이 수록된 학술논문까지 별도로 검색해야만 그 생명체에 관한 포괄적이고도 정확한 정보를 획득하여 연구를 수행할 수 있다.

이러한 정보검색 과정은 번거로울 뿐만 아니라 각 정보의 형태의 다양화로 인하여 원하는 정보를 한번에 검색하여 획득하는 것은 쉬운 일이 아니다. 그렇지만 다양한 형태의 다양한 생명과학 관련 정보를 통합함으로 연구자들이 원하는 값을 식별하고 얻을 수 있다면 생

명과학 연구의 효율성을 증진시킬 수 있다.

이런 이종의 데이터베이스들의 통합검색을 위한 시도 및 새로운 방법론은 지속적으로 개발 및 발전되고 있으며 생물학적 데이터베이스를 통합하여 검색을 제공하는 일반적인 방법으로는 첫째, 인덱스된 데이터 소스(Indexed data sources) 기법이 있다. 이 기법은 많은 수의 데이터 소스들을 인덱스하고 링크하는 것이다. 유저는 하나의 데이터 소스에 질의를 하고 관련된 다른 소스들의 정보의 링크를 따라간다. 생물학적 데이터베이스를 위한 키워드 인덱스 검색시스템으로 Sequence Retrieval System (SRS)이 있다. 둘째, 연합 데이터베이스(Federated databases) 기법이 있다. 이 기법은 각각의 소스의 데이터베이스들 안에 정보가 거주하며 Federated 시스템은 이종의 소스 데이터 스키마를 목적 스키마로 통합하기 위하여 common data model을 유지한다. 펜실베이니아 대학에서 개발한 K2/Kleisli와 IBM에서 개발한 Discovery Link를 예로 들 수 있다. 셋째, 스키마 통합(Schema Integration) 기법이 있다. 스키마 통합의 목표는 강요나 방해 없이 통합된 query 인터페이스를 다수의 자율적이고 이종의 데이터 소스로 제공하는 것을 목표로 하며, 유저들은 Mediated 스키마의 관점으로 질의한다. 이 방법은 관계형 데이터베이스(RDBMS)에 가장 효율적이다.

위 방법론 중에 인덱스를 이용한 방법은 검색속도 측면에서 비효율적이고, 연합 데이터베이스의 경우에는 지역(하위) 시스템들이 구조적으로는 동질이지만 의미가 다른 데이터를 제공할 수 없다. 또 많은 데이터베이스가 연합에 참가하는 경우 각 하위시스템들은 그들 자료의 의미와 타 하위시스템의 의미를 중재해 주는 의미해석장치(semantic translator)를 장착하여야 하는 단점을 가지고 있다. 그러므로 최소한의 비용으로 효율적인 통합검색을 가능하게 하기 위하여 스키마 통합 기법을 기반으로 하는 것이 효율적일 것이다.

국내·외에서 생명정보 제공 서비스를 위한 웹사이트로는 바이오인포매틱스센터(http://www.ccbb.re.kr), 생물학연구정보센터(http://bric.postech.ac.kr), 미국 국립보건원 산하 생명공학연구소(NCBI, National Center for Biotechnology Information) (http://www.ncbi.nlm.nih.gov), 유럽 생물정보학연구실(EBI, European Bioinformatics Institute) (www.ebi.ac.uk) 등이 있지만 생명정보, 생물종정보, 문헌정보 등의 전반적인 통합검색을 제공하는 서비스는 아직 없는 실정이다.

이에 본 논문에서는 현재 한국과학기술정보연구원(KISTI)에서 서비스 중인 문헌정보시스템(NDSL), 생물다양성정보시스템(KBIF), 생명정보시스템(CCBB)에서 각각의 데이터베이스의 스키마를 분석하여 메타데이터를 추출하고, Mediated 스키마를 설계하여, 이종의 생명정보 관련 데이터베이스들을 통합검색 할 수 있는 스키마 통합 기반 생명정보 검색시스템(BIRS, Biological Information Retrieval System)을 설계하였다.

2. 이론적 배경

2.1 선행연구

대량의 데이터를 처리하고 관리하기 위한 정보수집, 분석, 가공, 구축, 검색 기법 및 기술에 관한 연구는 이미 수준급이며, 이를 이용하여 생명정보를 통합 검색하는 연구는 인간 유전자 지도의 완성과 생명과학의 발달에 힘입어 앞으로도 많은 연구가 진행될 것으로 보인다.

J. Leon Zhao(1997)는 다양한 조직의 환경에 관련된 질의 처리를 위한 새로운 방법인 스키마 조정(coordination)의 접근을 소개하였다. 이것은 논리적 데이터베이스 구조와 데이터 영역 그리고 제약사항의 대립 없이 데이터베이스 통합을 달성한다.

Baker, P 등(1998)은 TAMBIS를 설계하였는데, 이것은 분자생물학과 생명정보학을 위한 도메인 온톨로지이며, 검색 기반 정보통합 시스템이다. 이러한 질의는 독립적인 질의의 모음으로 다시 작성되며, 미들웨어 계층에서 실행된다. 이 논문은 시스템안의 온톨로지에 의하여 수행되는 점으로, 다른 프로젝트와 구별된다.

윤홍원(2000)은 분산된 생물정보 데이터베이스의 통합검색시스템에 관한 연구를 수행하였다. 염기 서열, 단백질 서열, 유전자 서열 등의 서열 데이터베이스와 단백질 3차 구조를 제공하는 구조 데이터베이스 등 전 세계적으로 분산되어 있는 다양한 생물정보 데이터베이스의 효율적인 검색시스템인 GenPlus를 제안하였다. 제안한 GenPlus에서는 12개의 생물정보 데이터베이스를 질의 파서, 지역 데이터베이스, 질의생성기, 결과 처리기로 구성하였고, 서열과 키워드를 이용한 서열, 구조, 유전자 및 문헌정보의 통합검색을 제공하도록 설계하였다.

신종환 등(2002)은 XML 기반의 생물정보학시스템 프레임워크를 제시하였다. 생물정보학에서의 통합의 필요성을 기술하고 통합에 관련된 요구사항에 맞도록 BIS(Bioinformatics Information System)를 데이터 모델과 기업측면의 전체 BIS 아키텍처 수준으로 나누어 XML 기반의 통합 BIS 프레임워크를 제시하고 구성요소들을 정의하였다. 데이터 모델 수준에서의 통합은 염기와 단백질 데이터, 단백질의 2차 및 3차 구조 그리고 명명법과 같은 표준으로 표현하고 교환할 것인가를 다루었으며, 기업측면에서는 BIS 아키텍처 수준에서의 통합은 염기 혹은 단백질 서열정보와 각종 생명정보학 도구 등을 포함한 대규모의 데이터베이스들을 통합하고 관리하는 문제들을 포함한다.

이희전 등(2003)은 웹서비스 기반 유전자 주석정보 통합검색시스템을 구축하였다. 유전자 서열 주석정보들의 통합검색을 위하여 웹서비스 기술을 기반으로 분산된 데이터베이스 중 특히 유전자 주석 데이터와 관련된 기존 데이터베이스들로부터 통합검색시스템을 구축하

였다. 본 논문에서는 BioDAS의 웹서비스 개념을 이용, 분산된 주석 데이터 서버들 간의 통합검색시스템을 구축함으로써 메타검색시스템을 구현하였다. 본 시스템은 사용자에게 메타검색 기능 및 결과 저장기능을 제공해 주며 외부 사용자에게 웹서비스를 제공한다.

이수정 등(2004)은 바이오 관련 데이터베이스의 통합과 연동기능의 제공의 필요성을 기술하고 현재까지 진행되고 있는 많은 통합 연구 시스템의 대부분이 링크 기반, 데이터웨어하우징 구축 기반으로 하고 있어서, 데이터 스키마나 데이터의 변경 시, 실시간 업데이트와 같은 문제점을 가지고 있기에, 이러한 비효율적인 면을 개선시키고자, 플랫폼과 스키마의 변화에 구애 받지 않고 서비스를 가능하게 하는 웹서비스를 이용한 바이오 서열 데이터의 데이터베이스와 통합검색시스템을 개발하였다.

Jim Gray 등(2005)은 smart notebook을 사용한 데이터와 데이터 분석 계층에 관한 내용을 다루고 있다. smart notebook의 방대한 데이터 저장소에 있는 전산화한 자원들을 탐험할 수 있는 분석도구를 통하여 데이터에 접근할 수 있으며, smart notebook의 목표는 과학자들이 세계적으로 분포되어 있는 데이터를 탐험할 수 있게 해주는 것이다.

Shuai 등(2006)은 지리정보시스템을 이용하여 캐나다에서 웨스트 나일 바이러스(West Nile virus)로 죽은 조류를 실시간 감시하기 위한 파일롯 시스템을 개발하였다. 통합된 실시간 감시 시스템은 전통적인 실시간 웹 기반의 감시 컴포넌트, 통합된 실시간 GIS 컴포넌트와 통합된 Open GIS 컴포넌트를 포함하고, 새로이 개발된 웹 GIS 기술에 의하여 구동되고 링크된다.

2.2 관련 정보 현황

본 논문에서 연계 및 통합 검색의 대상이 되는 데이터베이스는 유전자/단백질 서열정보, 생물다양성정보와 생명체 분포를 표현하는 지리정보, 그리고 관련 논문정보이며, 그 내용을 소개하면 다음과 같다.

2.2.1 생명정보(바이오인포매틱스) 현황

생명정보 데이터의 분석과 저장 관리를 위해서는 데이터의 특성을 이해하고 특성에 맞는 분석모델의 선택과 이에 대한 제약사항을 추가해야 한다. 특히 HGP(Human Genome Project) 이후 출현한 생명정보 데이터는 그 종류가 다양하고 항상 변화되고 유동적이므로 더욱 그 특성에 대한 분석이 중요하다.

1) 유전자(gene) 정보

통상적으로 유전자란 동물의 핵 속에 들어있는 유전자 전체를 말하며, 염색체의 일부인 DNA 절편에 존재하는 기능단위라고 이해하고 있다. 유전체(genomic)란 gene(유전자)+some(항체)의 합성어로 주로 유전자 지도 작성, 돌연변이의 생산 및 분석, 질병유전자 및 특수형질과 관련한 원인 유전자규명, 분자표

지인자(DNA marker)개발 연구 등을 포함한 정보를 말한다. 대표적인 유전자 데이터베이스로는 Genbank가 있다.

2) 단백질(protein) 정보

단백질은 20여 개의 아미노산을 구성단위로 하여 조성되어 있으며, 생체를 구성하고 유지하고 제어하는 물질이다. 세포내에서 합성되는 단백질을 조사하기 위해서는 먼저 각 세포내에서 합성되는 모든 종류의 단백질을 분리할 필요가 있다. 전하의 크기에 따라 단백질을 분리하는 전기영동법과 질량에 따라서 단백질을 분리하는 mass-spectrum 방법을 이용하여 2차원 상에서 다양한 단백질 분리 패턴을 얻게 되며, 이를 분석하여 각 세포내에서 합성된 단백질 전체를 파악할 수 있다. 대표적인 단백질 데이터베이스는 PDB, PIR, Swiss-prot 등이 있다.

2.2.2 생물다양성정보 현황

1) 생물다양성정보기구

생물다양성정보에 관한 국제기구로는 세계생물다양성정보기구(GBIF)가 있다. GBIF는 지구상에 존재하는 생명체에 관한 생물다양성정보를 범세계적인 네트워크로 생물다양성의 보전과 이용, 공평한 공유를 목적으로 1999년 설립되었으며 GBIF 포털, 국가거점노드, 데이터노드로 구성되었다.

2) 생물다양성정보 표준

생물다양성정보는 전 세계적으로 각 기관의 특성에 맞게 콘텐트, 스키마, 구조 등이 결정되어 데이터베이스에 저장되어 서비스되고 있다. 생물개체의 공통된 특성을 추출하여 생물다양성 데이터로의 접근을 쉽게 할 수 있도록 만든 대표적인 데이터 표준으로는 Darwin-Core와 ABCD 형식이 있다.

2.2.3 지리정보 현황

지리정보시스템(GIS, Geographic Information System)이란 인간생활에 필요한 지리정보를 효율적으로 활용하기 위한 정보시스템의 하나이다. 지리정보시스템의 대표적인 소프트웨어로는 ArcGIS, Google Maps, World Wind 등이 있다.

1) ArcGIS

ArcGIS는 미국 ESRI사에서 개발한 지리정보분석 소프트웨어이며, 세계적으로 가장 많이 이용되고 있는 GIS 프로그램 중 하나이다. ArcGIS는 다양한 공간분석과 표현기능을 갖고 있어 현재 도시계획, 토목, 건축, 조경 등과 같은 공학 분야를 비롯하여 사회과학 분야 및 국방, 행정, 교육 등 다양한 분야에 이용되고 있다.

2) 구글 맵스(Google Maps)

구글 맵스는 전 세계에서 가장 많이 사용되고 있는 무료 지도서비스로 전 세계 160여 개국에 대한 상세한 지도 데이터 및 고해상도 이미지를 제공하고 있다. 구글 맵스는 정

밀한 도로정보와 한글 주소검색 기능을 제공하고 위성지도 및 지형정보 역시 함께 제공한다.

3) 월드 윈드(World Wind)

월드 윈드는 미국 NASA에서 2004년부터 개발하고 있는 삼차원 형태로 위성 영상을 볼 수 있는 프로그램이다. 위성 영상, 항공사진, 삼차원 형태의 GIS 데이터를 표현할 수 있다. 현재 지구뿐만 아니라, 달, 화성, 금성, 목성의 영상을 볼 수 있다. 월드 윈드는 OGC의 WMS(Web Map Service) 서버로부터 지형 데이터 및 지도를 다운받거나, ESRI Shapefile이나 KML 같이 자주 사용되는 형식으로 자료 표출이 가능하다.

2.2.4 학술논문정보 현황

1) PubMed

PubMed는 미국국립의학도서관(NLM)의 문헌 데이터를 검색하는 시스템으로, NCBI에서 무료로 제공하고 있다. 의학과 생명과학 분야의 문헌(1960년대 중기 이후)을 망라하고 있으며, 키워드를 검색창에 입력하면 간단히 문헌검색을 할 수 있다. 이미 보고된 문헌의 제목, 저자, 초록 외에 문헌에 따라서는 무료로 전문(full text)을 참조할 수 있으며, 검색결과 화면으로 직접 이동하기 위한 URL과 검색결과를 파일에 저장하는 등 결과를 이용자가 원하는 형식으로 제공받을 수 있다.

2) PubMed Central

PubMed Central은 자유롭게 웹에서 의학 논문을 누구나 볼 수 있게 수집하여 제공하는 사이트이다. 미국립의학도서관(NLM)에서 2001년부터 작업을 시작하였으며 PubMed보다 학술지 종수는 훨씬 적지만 이용자들이 전문을 무료로 제공받을 수 있다는 점이 PubMed Central의 가장 큰 장점이다. 그런 이유로 의학과 생명과학 연구자들은 PubMed보다 PubMed Central을 먼저 검색하는 경우가 늘어나고 있으며, 앞으로 PubMed Central의 영향력은 커지게 될 것으로 예상된다.

3) NDSL

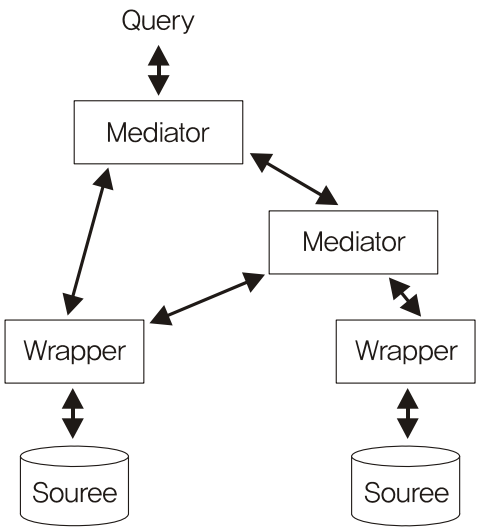
NDSL은 국내 학계, 연구계, 산업계의 모든 연구자를 위한 해외 학술 저널 및 프로시딩 포털사이트로서 5만6,000여 종의 학술저널과 18만8,000여 종의 프로시딩에 수록된 4200만 건 이상의 논문에 대한 서지-초록-원문 연계 One-Stop 서비스를 제공하고 학계, 연구계, 산업계, 의료계 등의 고급 연구인력 학술 전자정보를 최소한의 비용으로 자유롭게 활용할 수 있는 국가적 이용기반을 구축하고 있다.

3. 시스템 구조 설계

본 논문에서 설계한 시스템은 특성이 서로 다른 네 개의 시스템을 통합 및 연계하여 검

색할 수 있도록 하기 위하여 각 시스템마다 Mediated 스키마를 구축하여 연계되도록 하였다. 데이터 매핑(data mapping)은 두 개의 구분된 데이터 모델 사이의 엘리먼트(element)를 매핑하는 과정이다. 데이터 매핑은 상이한 용도의 데이터 소스 사이의 변환이나 중재, 데이터 계통 분석의 목적으로 데이터의 관계를 식별, 다중의 데이터베이스에서 하나의 데이터베이스로 데이터의 합병과 제거를 위하여 불필요한 컬럼(column)을 식별 등의 목적을 달성하기 위하여 사용된다. 즉, Mediated 스키마를 설계하기 위하여 메타데이터를 추출 및 분석하여 매핑하는 과정이 우선되어야 할 것이다.

생명정보 검색시스템의 연계는 〈그림 1〉에서 보는 바와 같이 Mediator를 사용하여 서로 다른 데이터베이스 스키마를 통합하도록 구성



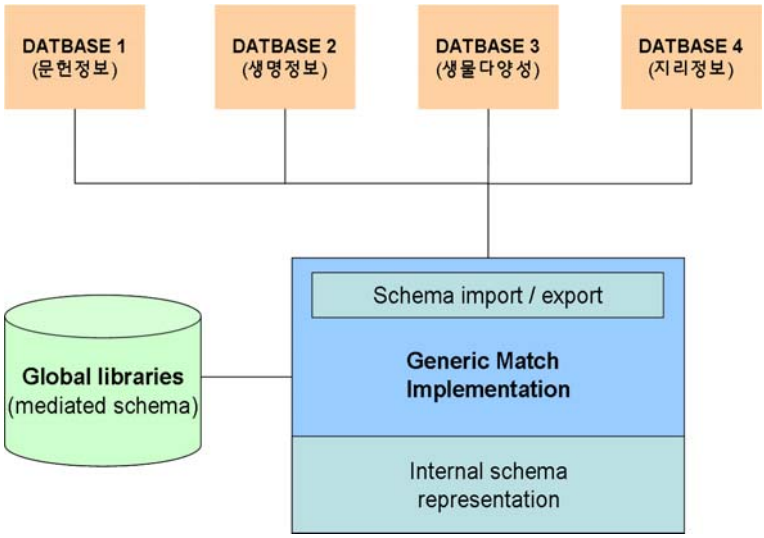
〈그림 1〉 Source와 Mediator의 관계도

하였다. 각각의 데이터베이스는 다른 개발자와 관점에 의해 개발되었기 때문에, 서로 다른 구조와 용어(terminology)를 가진다. 특히, 서로 다른 분야를 포함하는 데이터베이스의 통합을 시도할 때 어떤 문제가 발생할 수도 있다. 이를 위해서 각 데이터베이스의 스키마를 추출하고, Mediated 스키마를 구성하였으며 각 데이터베이스에 통합된 질의를 전송 가능하도록 본 시스템을 설계하였다.

Mediated 스키마는 각각의 데이터베이스 사이에 위치하며, 한 데이터베이스에서 다른 데이터베이스로 통합된 질의가 전달될 수 있도록 한다. Wrapper는 통합된 질의를 독립적인 데이터베이스가 이해 가능한 질의로 번역하여 전송한다. 이렇게 만들어진 질의는 각 데이터베이스의 소스에 전달되어 검색을 실행하게 되고, 이러한 과정을 거친 정보검색 결과는 다시 웹브라우저로 전달된다.

웹브라우저에서는 생명체의 명칭으로 검색을 시작하여, 다양한 다른 시스템의 정보를 하나의 화면에서 볼 수 있도록 설계하였다. 즉, 지리정보시스템을 통한 생물의 위치정보와 생물다양성 정보시스템을 통한 생물의 특징 및 생명정보시스템을 통한 유전자 서열정보를 하나의 통합된 질의로 검색 가능하다.

〈그림 2〉는 Global libraries에 속하는 Mediated 스키마가 다수의 자율적인 데이터베이스의 스키마를 통하여, 각 데이터베이스로 연동되는 흐름을 보여주는 구조도이다. 이러한 시스템을 구현하기 위해서 우선 스키마



〈그림 2〉 스키마 연동 구조도

의 특성을 파악하고 관계를 식별해야 할 것이다. 그리고 일관된 관점으로 Mediated 스키마를 작성해야 할 것이다.

4. 메타데이터 스키마 설계

4.1 메타데이터 추출

통합검색을 위해서 다음과 같이 생명정보, 생물다양성, 지리정보, 문헌정보 데이터베이스를 선정하였으며 특징은 〈표 1〉과 같다. 각각의 데이터베이스가 해당 분야에서 가장 대중적이며 주요한 역할을 수행하고 있다고 판단되기 때문에, 이와 같이 데이터베이스를 선정하였다. 여기서 Genbank는 CCBB에서 kristal 검색엔진을 통하여 서비스되고 있는 것을 말한다.

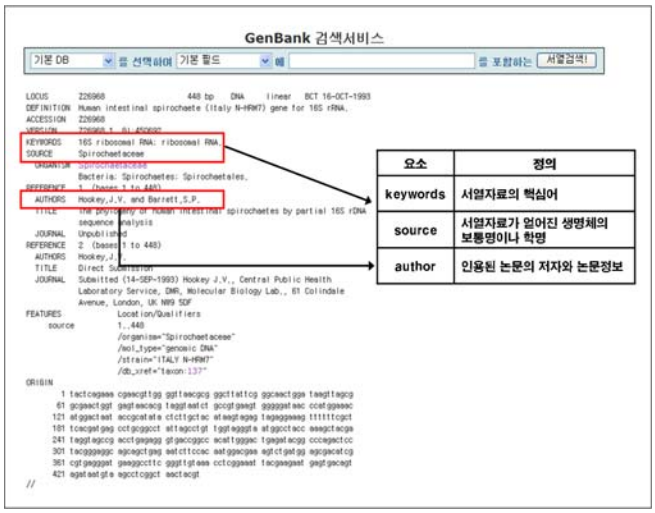
스키마를 통합하기 위한 첫 번째 단계는 각각 데이터베이스의 내부스키마의 특징을 식별하는 것이며, 이것을 통하여 생명정보시스템(CCBB,

〈표 1〉 데이터베이스의 특징

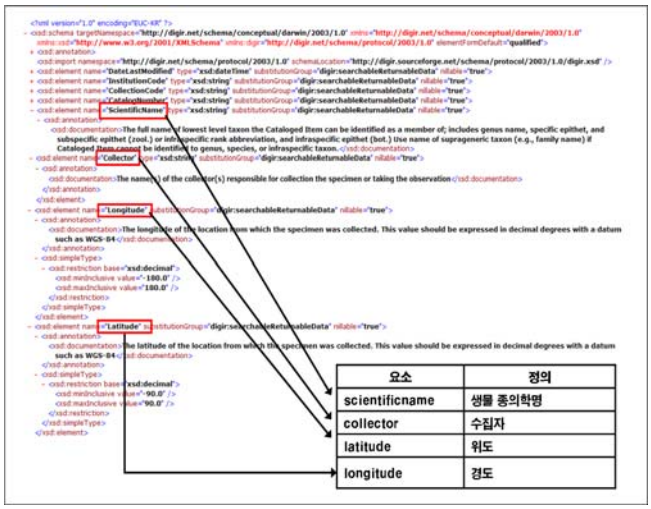
데이터베이스	특 징
Genbank	세계 최대의 공개된 유전정보 데이터베이스
KBIF	세계생물다양성정보기구(GBIF)의 한국거점노드
Google Maps	전 세계에서 가장 많이 사용되는 무료지도 서비스
NDSL	국내 및 해외 학술 저널 및 프로시딩 포털사이트

Genbank), 생물다양성 정보시스템(KBIF), 지리정보시스템(Google Maps), 과학기술정보시스템(NDSL)의 각 스키마에서 주요 메타데이터 요소를 추출할 것이다. 한번 식별이 완료되면 일관적인 관점으로 Mediated 스키마로 단합될 수 있을 것이며, 차후 생성된 질의는 오리지널 스키마로부터 통합된 표현으로 데이터의 변환이 허용될 것이다.

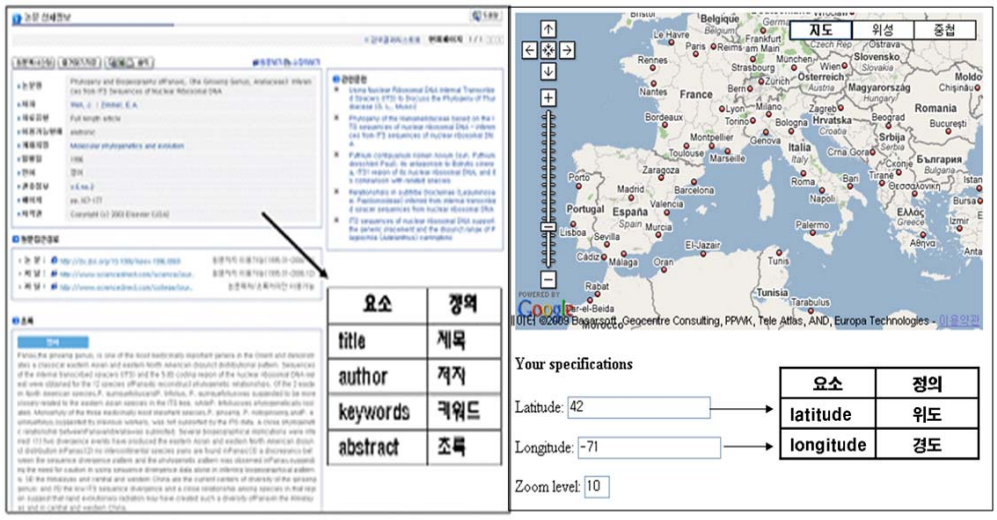
〈그림 3〉은 Genbank 요소에서 keywords, source, authors를 추출하여 작성한 메타데이터이며, 〈그림 4〉는 KBIF 스키마 요소 중에서 scientificname, collector, latitude,



〈그림 3〉 Genbank 데이터 요소와 추출된 메타데이터



〈그림 4〉 생물다양성정보 데이터요소와 추출된 메타데이터



〈그림 5〉 논문정보, 지리정보 데이터요소와 추출된 메타데이터

longitude 등을 추출하여 작성한 메타데이터 스키마이다.

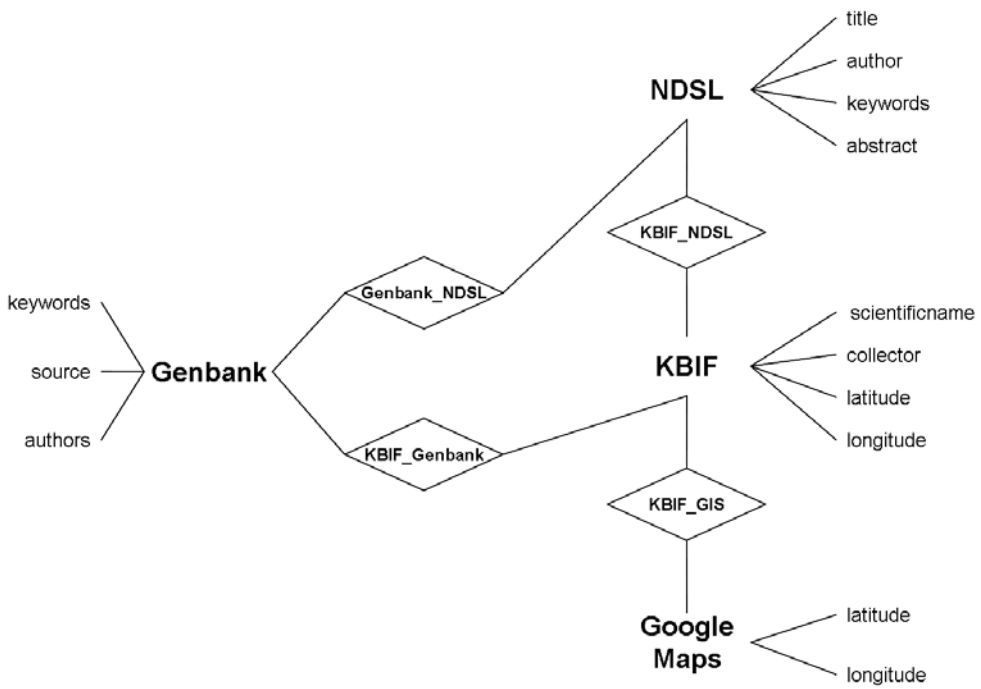
과학기술정보시스템의 논문정보 메타데이터 요소 중에서는 〈그림 5〉와 같이 title, author, keywords, abstract 등을 추출하였으며, 지리정보 메타데이터는 Google Maps에서 좌표를 측정할 수 있는 latitude, longitude를 추출하였다.

4.2 Mediated 스키마 설계

생물다양성정보(KBIF), 생명정보(Genbank), 논문정보(NDSL), 지리정보(GIS) 등 각각의 스키마에서 추출한 메타데이터 요소는 〈표 2〉과 같으며, 각 데이터베이스 사이의 Mediated 스키마는 〈그림 6〉과 같은 관계를 가질 것이다. 즉, Genbank_NDSL은 Genbank와 NDSL 사이의 Mediated 스키마를 의미한다.

〈표 2〉 추출된 메타데이터 요소

데이터베이스명	메타데이터 요소
Genbank	keywords, source, authors
KBIF	scientificname, collector, latitude, longitude
NDSL	title, author, keywords, abstract
Google Maps	latitude, longitude

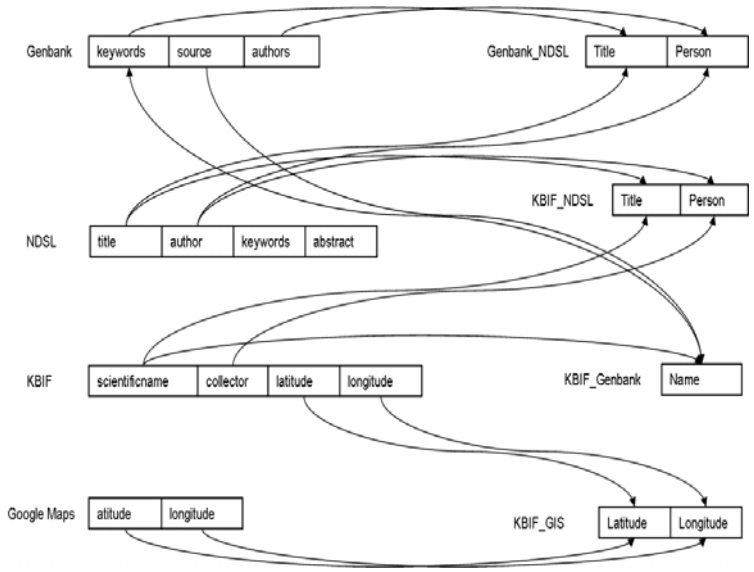


〈그림 6〉 Mediated 스키마 관계도

각각의 Mediated 스키마를 작성하기 위하여, 공통속성을 가질 Title, Person, Name, Position을 선정하여, 생물다양성정보(KBIF), 생명정보(Genbank), 논문정보(NDSL), 지리정보(GIS)에서 추출한 메타데이터를 〈표 3〉과 같이 상관관계를 정의하였으며, 4개의 Mediated 스키마 Genbank_NDSL, KBIF_NDSL, KBIF_Genbank, KBIF_GIS와의 각 데이터베이스에서 추출된 메타데이터 요소와의 속성 관계는 〈그림 7〉과 같다.

〈표 3〉 공통 속성 정의

속성	KBIF	Genbank	NDSL	Google Maps
Title	scientificname	keywords	title	–
Person	collector	authors	author	–
Name	scientificname	source	title	–
Position	latitude	–	–	latitude
	longitude	–	–	longitude



〈그림 7〉 속성 관계도

5. 검색 인터페이스 설계

5.1 사용자 질의 구성

사용자 질의를 Mediated 스키마의 입장에서 재구성하는 것은 통합 질의가 각 데이터베이스의 스키마를 통하여 소스로 직접 접근하도록 한다는 것을 의미한다. 〈표 4〉와 같은 다음의 질의문은 〈표 3〉에서 정의한 공통 속성인 Title, Position을 선정하여 질의를 재구성하는 시나리오를 보여주고 있으며, 각각 속성의 연관관계는 〈그림 7〉과 같다. 마지막으로 위와 같이 재구성을 거친 질의는 〈표 5〉와 같은 질의 처리 단계를 거쳐서 각각의 데이터베이스로 전송된다.

〈표 4〉 질의문

q(Title, Position) : -	KBIF_NDSL (title, author), KBIF_Genbank (Name), KBIF_Google Maps (Position)
Q(Title, Position) : -	KBIF (scientificname, collector, latitude, longitude), NDSL (title, author, keywords, abstract), KBIF (scientificname, collector, latitude, longitude), Genbank (keywords, source, authors), KBIF (scientificname, Collector, latitude, longitude), Google Maps (latitude, longitude)

〈표 5〉 질의 처리 단계

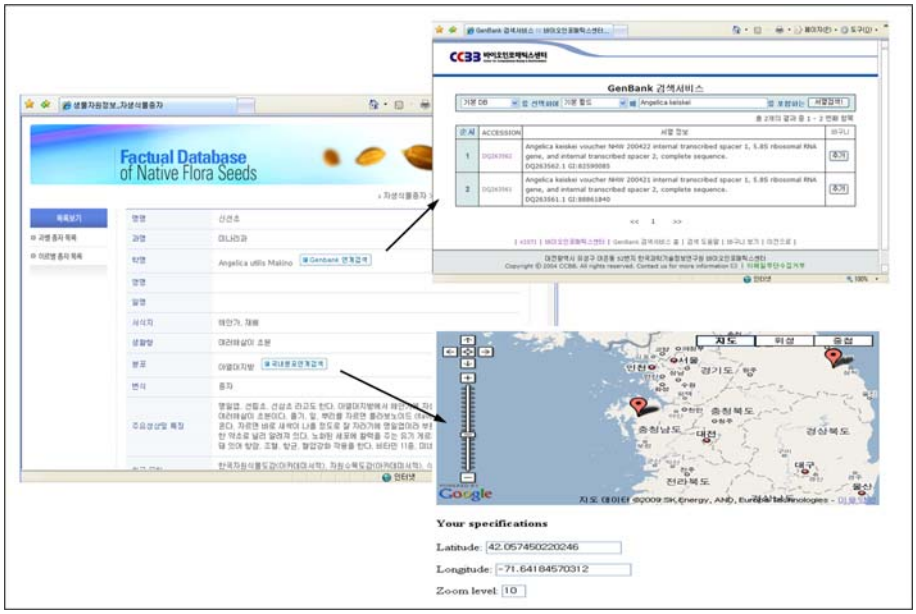
query issued	SELECT [Title], [Position] FROM KBIF, Genbank, NDSL WHERE KBIF.[Title]=(Genbank, NDSL).[Title]
subquery created	DATABASE KBIF, NDSL SELECT [Title] FROM WHERE KBIF.[Title]=NDSL.[Title]
	DATABASE KBIF, Genbank SELECT [Title] FROM WHERE KBIF.[Title]=Genbank.[keywords]
	DATABASE KBIF, Google Maps SELECT [Position] FROM WHERE KBIF.[Position]=Genbank.[Position]
attributes converted	DATABASE KBIF, NDSL SELECT FROM WHERE KBIF.scientificname=NDSL.title
	DATABASE KBIF, Genbank SELECT FROM WHERE KBIF.scientificname=Genbank.keywords
	DATABASE KBIF, Google Maps SELECT FROM WHERE KBIF.latitude=Google Maps.latitude AND KBIF.longitude=Google Maps.longitude

5.2 BIRS 인터페이스 설계

본 논문에서 설계한 생명정보 검색시스템(BIRS, Biological Information Retrieval System)의 인터페이스는 일반적인 검색과 같이 사용자가 생명체에 대한 검색을 하면 Mediated 스키마를 통하여 각 데이터베이스로 통합검색을 하여 검색한 결과의 목록을 보여주고 목록 중 하나를 선택하면 원하는 데이터베이스로의 결과로 이동한다.

또한 BIRS 정보검색 인터페이스를 통하여 단순히 NDSL, KBIF, Genbank의 검색결과를 보여주는 것이 아닌, 각각의 데이터베이스들 간의 교차 연동이 가능하게 해 줄 것이다.

〈그림 8〉 BIRS 인터페이스 설계 화면



〈그림 9〉 생물다양성, 생명정보, 지리정보 연계검색 설계 화면

즉, 〈그림 9〉와 같이 통합검색 화면에서 선택한 생물다양성정보에서 생명정보의 연동이 가능하며, 생물다양성정보에서 분포정보로 연계된 필드를 선택했을 경우, 지리정보시스템에서 해당 생물의 분포에 관련된 위치정보를 보여준다.

6. 결론

본 논문에서는 생명체의 종정보, 유전자/단백질 정보, 위치정보, 논문정보 등을 통합검색이 가능하도록 스키마 통합을 이용하여 생명체를 검색하고 그 생명체가 분포하고 있는 위치정보와 생명체를 구성하고 있는 유전자 정보를 연계하여 서열정보를 볼 수 있을 뿐만 아니라 관련 논문까지 검색할 수 있는 생명정보 검색시스템 설계에 관하여 기술하였다.

본 논문에서 통합 및 연계 대상으로 삼은 시스템은 과학기술정보시스템(논문정보), 생물다양성 정보시스템(생물 종정보), 생명정보시스템(유전자정보), 지리정보시스템(생물 분포정보) 등 4개의 각기 다른 시스템이며, 이 4개의 시스템에서 스키마를 추출하여 Mediated 스키마를 생성하였다.

위에서 생성된 Mediated 스키마를 기반으로 생명체의 명칭(학명)을 이용하여 생명체에 대한 전반적인 정보와 지리정보시스템을 이용한 생물 분포정보를 시각적으로 확보할 수 있을 뿐만 아니라, 그 생물을 구성하고 있는 유전자/단백질 서열정보와 관련 문헌정보까지 연계하여 한 번의 키워드 검색으로 그 결과를 확인할 수 있도록 통합 및 연계 검색 처리 절차를 설계하였다.

이렇게 설계한 정보검색시스템은 semi-auto 기반의 스키마 통합 기술로서 관리자가 수동으로 독립적인 데이터베이스의 스키마들의 관계를 정의하고 매칭해야 하는 번거로움이 있을 수 있다. 그러나 시맨틱 웹으로 발전하기 위한 auto 스키마 통합 기술의 중간 단계로 본 시스템이 구현되어 생명과학 연구자들에게 제공된다면, 서로 다른 특성의 여러 개의 데이터베이스를 검색하기 위해 여러 사이트를 방문해야 하는 번거로움을 해결하고 동시에 다양한 검색결과를 보면서 분석하려는 생명과학 관련 연구자들의 연구 효율을 높이는 데 기여할 수 있을 것이다.

참고문헌

김대중, 박재홍, 안성수, 박형선. 2007. 매쉬업을 활용한 생물다양성 지리정보서비스. 『한국콘텐츠학회 2007 춘계 종합학술대회 논문집』, 5(1): 11-14.

김택천, 김석훈, 김진수. 2007. 유비쿼터스 환경을 위한 Web-GIS 기반의 객체 위치 정보 모니터링 시스템. 『한국해양정보통신학회논문지』, 11(9): 1755-1763.

이수정, 용환승. 2004. 웹서비스 기반 바이오 서열정보 데이터베이스 및 통합검색시스

템 개발.『한국정보처리학회논문지』, 11D (4): 755-764.

이희전, 용환승. 2003. 웹서비스 기반 유전자 주석정보 통합검색시스템 구축.『한국멀티미디어학회 2003년 추계학술발표대회 논문집(상)』, 2003: 355-358.

최요한, 유성준, 김민경, 박현석. 2004. 웹서비스 기반 바이오정보 통합 분석도구.『한국정보과학회 2004 봄 학술발표논문집』, 2004: 289-291.

Baker, P. G. Brass, A. Bechhofer, S. 1998. "TAMBIS: Transparent Access to Multiple Bioinformatics Information Sources." *Intelligent systems for mole-cular biology*, ISMB-98: 25-34.

BRIC-PDB 설명. [cited 2009. 01. 12]. 〈http://bric.postech.ac.kr/topic/75.htm〉, 〈http://bric.postech.ac.kr/info/review/20.html〉.

Davidson. S. B. Crabtree. J. Brunk, B. P. 2001. "K2/Kleisli and GUS: Experiments in integrated access to genomic data sources." *IBM systems journal*, 40(2): 512-531.

E. M. Zdobnov, R. Lopez, R. Apweiler. 2002. "The EBI SRS server-new features." *Bioinformatics*, 18: 1149-50.

Garcia-Molina, Hector, Papakonstantinou, Yannis, Quass, Dallan. 1997. "The TSIMMIS Approach to Mediation: Data Models and Languages." *Journal of intelligent information systems*, 8(2): 117-132.

Gray. J, Liu, D. T, Nieto-Santisteban, M. 2005. "Scientific Data Management in the Coming Decade." *SIGMOD record*, 34(4): 34-41.

Haas, L. M. Schwarz, P. M. Kodali, P. 2001. "DiscoveryLink: A system for integrated access to life sciences data sources." *IBM systems journal*, 40(2): 489-511.

J. Leon Zhao. 1997. "Schema coordination in federated database management: a comparison with schema integration." *Decision Support Systems*, 20(3): 243-257.

Jian shuai, Peter buck, Paul Sockett, Jeff Aramini, Frank Pollari. 2006. "A GIS-driven integrated real-time surveillance pilot system for national West Nile virus dead bird surveillance in Canada." *International Journal of Health Geographics*, 5: 1-17.

KISTI 과학기술정보통합서비스 웹사이트. [cited 2009. 01. 12]. 〈http://www.ndsl.kr/help_index.co〉.

KISTI 과학기술 학회마을 웹사이트. [cited 2009. 01. 12].

〈http://society.kisti.re.kr/main.html〉.

KISTI CCBB 웹사이트. [cited 2009. 01. 12]. 〈http://www.ccbb.re.kr/ccbb/abou tccbb/introduction.jsp〉.

Kristal. [cited 2009. 01. 12]. 〈http://www.kristalinfo.com/〉.

Rahm, Erhard, Bernstein, Philip A. 2001. "A survey of approaches to automatic schema matching." *The VLDB jour-nal*, 10(4): 334-350.

Shuai, Jiangping, Buck, Peter, Sockett, Paul. 2006. "A GIS-driven integrated real-time surveillance pilot system for national West Nile virus dead bird surveillance in Canada." *International journal of health geographics*, 5: 17.

WIKIPEDIA-Virtual glove. [cited 2008. 08. 04]. 〈http://en.wikipedia.org/wiki/Virt ual_globe〉.

Zhang, Bing. Kirov, Stefan. Snoddy, Jay. 2005. "WebGestalt: an integrated system for exploring gene sets in various biological contexts." *Nucleic acids research*, 33: 741-748.