GenBank를 활용한 이종의 콘텐츠 연계 프로토타입 시스템 개발 연구*

A Study on Development of GenBank-based Prototype System for Linking Heterogeneous Content

안 부 영** · 신 용 주*** · 김 대 환****

Bu-Young Ahn · Young-Ju Shin · Dea-Hwan Kim

차 례

1. 서론	4. 콘텐츠 연계 프로토타입 구현
2. GenBank 분석	5. 결론
3. 콘텐츠 연계 서비스 설계	· 참고문헌

초 록

생명정보 중에서 미국의 국립생명공학정보센터(NCBI)에서 제공하는 GenBank는 전 세계적으로 연구자들이 가장 많이 사용하는 대표적인 유전자정보 데이터베이스이다. 한국과학기술정보연구원(KISTI)은 GenBank의 최신 버전을 데이터베이스로 재구축하여 Bio-KRISTAL 검색엔진을 이용하여 국내 생명과학 연구자들에게 제공하고 있다. 본 논문에서는 GenBank 데이터베이스를 활용하여 과학기술정보 통합서비스인 NDSL의 논문정보, 특허정보, 생물다양성정보 등의 콘텐츠와 GenBank reference 필드와 organism 필드를 상호 연계하는 서비스 모델을 설계하고 프로토타입 시스템을 개발하였다. 이를 위하여 1) NCBI FTP 사이트에서 GenBank 데이터를 수집하여, 2) GenBank 텍스트 파일을 유전자 기본정보와 참고정보로 나누어 데이터베이스로 재구축하여, 3) GenBank reference 필드에서 논문 및 특허 정보 추출을 통한 새로운 테이블을 생성하여, 4) 데이터 맵핑 기술을 이용하여 GenBank 데이터와 NDSL 데이터가 상호 연계되어 서비스되는 프로토타입 시스템을 구현하여 이종의 콘텐츠간 연계 및 융합 서비스의 가능성을 확인하였다.

키 워 드

GenBank, 콘텐츠 연계, 콘텐츠 융합, 유전자정보, 논문서지정보, 특허정보, 생물다양성정보, 맵핑기술

- * 본 연구는 한국과학기술정보연구원 2008년 모험창의과제로 수행되었으며, 2008년 11월 “제30회 한국정보처리학회 추계학술대회”에서 발표한 설계에 관한 내용을 기초로 시스템을 구현하여 작성한 것임.
- ** 한국과학기술정보연구원 차세대연구환경개발실 선임기술원
(Senior Researcher, Dept. of Cyber Environment Development, KISTI, ahnyoung@kisti.re.kr)
- *** 한국과학기술정보연구원 지식기반실 선임연구원
(Senior Researcher, Dept. of Knowledge Resources, KISTI, yjshin@kisti.re.kr)
- **** 한국과학기술정보연구원 지식기반실 초청선임연구원
(Invited Senior Researcher, Dept. of Knowledge Resources, KISTI, dhkim@kisti.re.kr)
- 논문접수일자 : 2009년 10월 6일
 - 최종심사일자 : 2009년 11월 10일
 - 게재확정일자 : 2009년 11월 16일

ABSTRACT

Among biological information, GenBank, provided by the National Center for Biotechnology Information (NCBI) of the United States, is a representative database on genetic information and is the most widely used by researchers around the world. Korea Institute of Science and Technology Information (KISTI) visits NCBI on a regular basis and downloads the latest version of GenBank to reorganize the information gathered there into a database. This database is provided for Korean researchers of science and technology through the Bio-KRISTAL search engine, developed by KISTI. This study aims to design a service model that links information on papers, patents, and biodiversity and other contents of NDSL, an integrated service on scientific and technological information run by KISTI, with GenBank's reference and organism fields and to develop a prototype system. For this purpose, this paper explores the possibility of a linkage and convergence service between heterogeneous content by: (a) collecting GenBank data from NCBI's FTP site; (b) dividing GenBank text files into basic and reference genetic information and restructuring them into a database; (c) extracting article and patent information from the GenBank reference fields to generate new tables; and (d) leveraging data mapping technology to implement a prototype system where GenBank and NDSL data are interlinked and provided.

KEYWORDS

GenBank, Content Linking, Content Convergence, Genetic Information, Bibliographic Information, Patent Information, Biodiversity Information, Mapping Technology

1. 서론

생명과학 연구자뿐 아니라 분야별 전문 연구자들은 산재한 다양한 콘텐츠의 융합된 정보를 요구하고 있으며 그 요구를 충족시킬 수 있는 전문적인 콘텐츠 및 서비스 필요성이 대두되고 있다. 그러나 생명과학, 화학, 물리 분야 등 각 전문정보 간의 연계 서비스는 되어 있으나 전문정보와 학술논문과 연계되어 정보서비

스를 제공하는 사례는 거의 없는 실정이다.

전문정보와 학술논문의 연계 서비스에 관한 국내외 현황은 다음과 같다.

특허청의 전통지식포털(<http://www.koreantk.com>)에서는 국내 학술논문과 특허, 한의학의 약재, 처방, 약재의 성분 화합물 등이 서로 연계되어 서비스되고 있다. 한의학 관련 전문정보가 필요한 이용자들이 사이트에 접속하면 우리나라의 전통지식에 관한 전문적인 지

식과 더불어 학술논문을 동시에 검색이 가능한 사이트이다. 이렇게 앞선 서비스를 제공하는 것은 바람직한 현상이라 할 수 있다.

한국과학기술정보연구원(KISTI)의 화합물 DB(<http://chemdb.kisti.re.kr/>)는 화합물 정보와 해외 화합물구조, 스펙트럼, 독성 및 환경 정보 등을 연계하여 서비스하고 있다. 그러나 관련 학술논문과는 연계되어 있지 않다. 학술논문의 연계까지 제공한다면 관련 연구자들에게 깊이있는 서비스를 제공하게 될 것이다.

한국과학기술정보연구원(KISTI)의 무기결정구조 DB(<http://icsd.kisti.re.kr/>)와 플라즈마 물성 DB(<http://plasma.kisti.re.kr/>)는 학술문헌으로부터 색인된 물성정보로 데이터베이스를 구축한 경우로써 NDSL 학술논문과 연계 서비스를 제공하고 있다. 무기결정구조와 플라즈마 물성 정보를 필요로 하는 연구자들에게 좋은 반응을 얻고 있다.

미국립의학도서관(NLM, US National Library of Medicine) 국립생물공학정보센터(NCBI, National Center for Biotechnology Information)의 PubMed는 1949년부터 지금까지 발간된 4,800여종의 생명과학 및 의학 분야 저널에 수록된 약 1,787만1,483건(2008.3.31 현재)의 논문에 관한 서지사항 및 초록을 제공하고 있다. 그러나 관련 생명정보 및 생물다양성정보와의 연계 서비스는 제공되고 있지 않다.

NCBI의 유전자 데이터베이스인 GenBank

에서는 reference 필드를 지정하여 규명된 유전자와 관련 논문의 저자, 논문 제목, 저널명 등의 간략정보를 제공하고 있다. Pubmed에 등록된 저널의 논문일 경우에는 Pubmed와 연계되어 논문을 찾아 볼 수 있지만 Pubmed에 등록되지 않은 저널의 논문일 경우에는 연계되어 있지 않다. 또한 유전자 관련 생명체의 학명과 NCBI Taxonomy Browser를 연계하여 분류학적인 생물체 종정보 확인이 가능하며 종정보에서 유전자지도 등을 볼 수 있도록 연계 서비스를 제공하고 있지만 생물체의 분포정보는 제공하고 있지 않다. 유럽의 EBI(European Bioinformatics Institute)와 일본의 DDBJ(DNA Data Bank of Japan)에서는 GenBank의 실시간 미러사이트를 운영하고 있기에 NCBI와 동일한 서비스를 제공하고 있다. 관련 정보를 다양하게 이용 가능하도록 정보서비스를 제공한다면 연구자들의 연구효율을 높이는데 기여할 수 있으리라 사료된다.

이에, KISTI에서 기 구축한 과학기술정보 통합서비스(NDSL)의 학술논문정보, 특허정보, 생물다양성정보와 GenBank를 연계함으로써 데이터의 부가가치 향상 및 고급 학술정보 서비스 필요를 충족시킬 필요성이 있다. CCBB 웹사이트에서 제공하고 있는 GenBank를 활용하여 학술논문정보, 유전자정보, 생물다양성정보, 특허정보를 상호 연계할 수 있는 서비스 방안을 제안하고자 한다. 이를 위해 본 논문에서는 <그림 1>과 같이 GenBank의 reference 필드와 source 필드에서 논문정보(논문

〈그림 1〉 이종의 콘텐츠 연계 정보 서비스 모델

제목, 저자, 수록처), 특허정보(특허명칭, 특허번호), 생물다양성정보(학명, 보통명)를 분석 및 추출하여 KISTI에서 구축하여 운영하는 과학기술정보 통합서비스인 NDSL(<http://ndsl.kr>)의 학술논문정보, 특허정보, 생물다양성정보와 상호 연계 가능한 정보서비스 모델을 설계하고 설계된 내용을 기반으로 프로토타입 시스템을 개발하고자 한다.

2. GenBank 분석

2장에서는 본 논문에서 제안할 콘텐츠 연계

프로토타입 시스템에서 가장 기본이 되는 GenBank의 개요와 콘텐츠 연계에 주요하게 사용될 데이터 필드(reference, organism)에 관하여 분석하였다.

2.1 GenBank 개요

인간은 약 100조개의 세포로 구성되어 있으며, 세포 내에는 세포핵이 존재한다. 세포는 23쌍의 염색체로, 23쌍의 염색체는 31억 개의 염기쌍으로 구성되어 있으며, 염기는 시토신(C), 구아닌(G), 아데닌(A), 티민(T)으로 구성되어 있다. 이렇게 규명된 유전자 염기서열

```

LOCUS~~~~~XELSRCC2~~~~~115 bp~~~~~mRNA~~~~~linear~~~~~VRT
28-APR-1993
DEFINITION~~X.laevis~Rous~sarcoma~virus~transforming~protein~mRNA,~3' end,
ACCESSION~~~M30860
~~~~~VERSION~~~~M30860.1~~GI:214812
KEYWORDS~~~~transforming~protein
~~~~~SEGMENT~~~~2 of 2
SOURCE~~~~~Xenopus~laevis~(African~clawed~frog)
~~~~~ORGANISM~~~~Xenopus~laevis
~~~~~Eukaryota~~Metazoa~~Chordata~~Craniata~~Vertebrata~~
~~~~~Euteleostomi~~
~~~~~Amphibia~~Batrachia~~Anura~~Mesobatrachia~~Pipoidea~~
~~~~~Pipidae~~
~~~~~Xenopodinae~~Xenopus~~Xenopus
REFERENCE~~~~1~~(bases 1 to 115)
~~~~~AUTHORS~~~~Steele,R,E
~~~~~TITLE~~~~Two divergent cellular src genes are expressed in Xenopus
laevis
~~~~~JOURNAL~~~~Nucleic Acids Res. 13 (5): 1747-1761 (1985)
~~~~~MEDLINE~~~~85215578
~~~~~PUBMED~~~~2987836
COMMENT~~~~~Original source text: X.laevis (female) erythrocyte, cDNA to
mRNA
FEATURES~~~~~Location/Qualifiers
~~~~~source~~~~~1..115
~~~~~organism="Xenopus~laevis"
~~~~~mol_type="mRNA"
~~~~~db_xref="taxon:8355"
~~~~~CDS<1..69
~~~~~note="transforming~protein~(src)"
~~~~~codon_start=1
~~~~~protein_id="AAA49965.1"
~~~~~db_xref="GI:214815"
~~~~~translation="LQAFLEDYFTATEPQYQPGDNL"
ORIGIN~~~~~Undetermined number of bp after segment 1.
~~~~~1~ctgcaggcgt~tctcgaggac~ctatttaca~gctaccgaac~cgcagtagca~
gacctggggac
~~~~~61~aacctttagc~cttcgctcat~aatcaagaga~catgtatagg~actcttagga~aacag
//

```

〈그림 2〉 GenBank 검색화면

을 데이터베이스로 구축하여 인간의 질병연구 및 치료에 활용하고 있다. 이와 같은 유전자 데이터베이스 중에서 전 세계적으로 가장 많이 사용되는 것은 미국 국립보건원(NIH, National Institutes of Health)의 국립생물 공학정보센터(NCBI, National Center for Biotechnology Information)에서 운영하는 GenBank이다.

GenBank는 염기서열 데이터베이스로 세계 각지에서 연구자들이 각자 등록한 서열 데이터를 다양한 각도의 분석 결과와 함께 제공한다. GenBank와 실시간으로 데이터 미러를 하는

기관은 유럽의 유럽분자생물학실험실(EMBL, European Molecular Biology Laboratory)과 일본의 DNA데이터뱅크(DDBJ, DNA Data Bank of Japan)가 있다.

NCBI는 또한 생물, 의학 분야 최대 문헌정보서비스인 Pubmed를 운영하고 있기에 〈그림 2〉에서 보는 바와 같이 GenBank reference 필드에 Pubmed id를 링크하여 서비스를 제공하고 있다. 그러나 Pubmed에 등재되지 않은 논문은 링크되어 있지 않아 Pubmed 이외의 논문을 필요로 하는 이용자들에게 불편함을 주고 있다.

2.2 GenBank 분석

NCBI, EMBL, DDBJ에서는 GenBank를 무상으로 다운로드 할 수 있도록 FTP 사이트를 운영하고 있다. 본 연구를 위해 GenBank release 163(2007년 12월, 8천4백만 건) 데이터를 다운로드 받아 분석하여 필요한 필드를 추출하였다. 참고로 압축 해제 시 GenBank 파일의 개수는 1,380개이며 파일 당 약 250MB, 약 400만 라인 정도로 구성되어 있었다.

2.2.1 reference 논문 필드 분석

GenBank 필드 중에서 reference 필드는 <표 1>과 같은 항목으로 구성되어 있으며 reference는 유전자정보 한 개당 N개까지 기술이 가능하다. 약 8천4백만 건의 유전자정보 전체 데이터의 reference 건수는 약 1억 건 정도였다. 이 결과로 GenBank 유전자정보 1건당 약 1.2건의 reference가 기술되어 있음을 알 수 있었다.

GenBank의 reference 필드를 추출하여

그 유형을 분석한 결과 <표 2>와 같이 정상적으로 필드가 기술된 경우, Unpublished인 경우, Direct submission인 경우, Patent인 경우 등 4가지 유형으로 나타났다.

GenBank는 텍스트 파일로 FTP 사이트를 통하여 130여개의 압축된 파일로 제공된다. 그래서 연계를 위하여 텍스트 파일의 압축을 풀어 유전자 기본정보와 reference 정보를 추출하여 MySQL 데이터베이스로 변환하였다. 변환 작업 결과 유전자정보 건수는 8,411만 2,248건, reference 건수는 1억416만7,000건으로 나타났으며 reference 유형별 데이터 분포는 <그림 3>과 같다. <그림 3>의 정상 논문 8,255만5,724건 중에서 Pubmed id를 가지고 있는 논문은 3,343만903건(40%)이고, Pubmed id를 가지고 있지 않은 논문은 4,912만4,821건(60%)이었다. 즉 GenBank 유전자정보의 reference-논문정보 중에서 60% 정도가 Pubmed와 연계되어 있지 않다는 것을 알 수 있었다(<그림 4> 참조).

<표 1> GenBank reference 필드 구성

필드명	데이터 기술 내용 예
REFERENCE	1 (bases 1 to 399)
AUTHORS	Belshaw,R., Fitton,M., Herniou,E., Gimeno,C. and Quicke,D.L.J.
TITLE	A phylogenetic reconstruction of the Ichneumonoidea (Hymenoptera) based on the D2 variable region of 28S ribosomal RNA
JOURNAL	Syst. Entomol. 23, 109-123 (1998)
MEDLINE	85215578
PUBMED	2987836

〈표 2〉 GenBank reference 필드 4가지 유형

유형	필드명	데이터 기술 내용 예
정상	AUTHORS	Harano, Y., Suzuki, I., Maeda, S., Kaneko, T., Tabata, S. and Omata, T.
	TITLE	Identification and nitrogen regulation of the cyanase gene from the cyanobacteria <i>Synechocystis</i> sp. strain PCC 6803 and <i>Synechococcus</i> sp. strain PCC 7942
	JOURNAL	J. Bacteriol. 179 (18), 5744-5750 (1997) 9294430
Unpublished	AUTHORS	Chen, W. and He, W. B.
	TITLE	Nucleotide Sequence and Characteristics of beta-amylase Gene from <i>Bacillus firmus</i>
	JOURNAL	Unpublished
Direct submission	AUTHORS	Iwabuchi, T.
	TITLE	Direct Submission
	JOURNAL	Submitted (25-DEC-1996) Tokuro Iwabuchi, Shiseido Research Center, Pharmaco Science Laboratories; 1050 Nippa, Kouhoku-ku, Yokohama, Kanagawa 223, Japan (E-mail: PEH01461@niftyserve.or.jp, Tel: +81-45-542-1337, Fax: +81-45-545-5931)
Patent	AUTHORS	Koizumi, S., Yonetani, Y. and Teshiba, S.
	TITLE	Process for producing riboflavin
	JOURNAL	Patent: US5589355-A 31-DEC-1996;

〈그림 3〉 reference 유형별 분포 현황

〈그림 4〉 Pubmed 연계 논문 현황

〈표 3〉 GenBank organism 필드 분석 결과

학명	건수	비율
homo sapiens	11,243,969	13.37%
mus musculus	7,274,491	8.65%
marine	5,731,044	6.81%
zea mays	2,878,332	3.42%
sus scrofa	2,121,442	2.52%
bos taurus	2,063,112	2.45%
arabidopsis thaliana	1,949,950	2.32%
danio rerio	1,591,010	1.89%
oryza sativa	1,587,998	1.89%
xenopus tropicalis	1,418,415	1.69%
rattus norvegicus	1,309,457	1.56%
unknown	1,267,148	1.51%
canis lupus familiaris	1,222,358	1.45%
triticum aestivum	1,104,170	1.31%
sorghum bicolor	1,011,769	1.20%
oryzias latipes	963,879	1.15%
gallus gallus	807,479	0.96%
glycine max	766,948	0.91%
drosophila melanogaster	740,537	0.88%
기재 안됨(NULL)	3,339,238	3.97%
ciona intestinalis의 생물종	33,719,502	40.09%
합계	84,112,248	100.00%

ganism 필드가 있다. 〈표 3〉에서 보는 바와 같이 organism 필드를 분석한 결과 인간에 관한 유전자가 1,124만3,969건으로 전체의 13.37%로 제일 많은 것으로 나타났고, 그 다음으로는 쥐에 관한 유전자가 727만4,471건으로 8.65%, 고래에 관한 유전자 등 약 20종의 생물종이 GenBank에 등록된 유전자 서열에서 60%정도를 차지하고 있었다. 그러나 333만 9,238건은 생물 종이 기재되어 있지 않았다.

3. 콘텐츠 연계 서비스 설계

지금까지 GenBank 데이터 중에서 reference 필드를 분석하여 그 결과를 알아보았다. 3장에서는 2장의 분석 결과를 바탕으로 기존 KISTI 콘텐츠(논문정보, 특허정보, 생물 다양성정보)와 GenBank를 상호 연계하기 위한 맵핑 테이블을 설계하였고 KISTI CCBB 웹사이트(<http://www.cccb.re.kr>)에서 서비

스 중인 GenBank 데이터베이스를 중심으로 연계 서비스 모델을 설계하였다.

생물다양성 종정보를 저장할 테이블(tb_genbank_taxonomy) 등 10개이다.

3.1 테이블 설계

본 연계 모형에서 콘텐츠간 연계를 위해 사용될 데이터베이스 테이블은 <그림 8>의 ERD(Entity Relationship Diagram)에서 보는 것과 같이 GenBank에서 추출한 유전자 기본 정보를 저장할 테이블(tb_GenBank), reference 필드를 저장할 테이블(tb_GenBank_ref), reference 필드의 특허정보를 저장할 테이블(tb_GenBank_patent), reference 필드의 논문 정보를 저장할 테이블(tb_GenBank_journal, tb_ndsl_artbrif), organism 필드에서 추출한

3.2 콘텐츠 연계 서비스 설계

3.2.1 NDSL 논문 콘텐츠

전체 논문 8,255만5,724건 중 Pubmed id 보유 3,343만903건을 제외한 pubmed id 미보유 논문에 대해 NDSL 논문과의 연계를 시도하였다. 이를 위한 Pubmed id 미보유 논문의 중복 제거 작업을 수행하면 서버의 메모리(4GB)를 초과하게 된다. 그래서 locus의 처음 2자리를 잘라서 영역을 나누고 200~400만 건 단위로 중복을 제거한 후 tb_GenBank_journal_distinct에 로딩하였으며, tb_gen-

<그림 8> ERD(콘텐츠간 매핑을 위한 테이블)

bank_journal_distinct의 데이터를 다시 중복 제거하여 tb_GenBank_journal_distinct1에 입력하는 절차를 반복 수행하였다.

그리고 tb_GenBank_journal_distinct1의 titlesort와 tb_ndsl_artbrif의 titlesort의 컬럼을 조인하여 NDSL ID(ARTCTRLNO)를 tb_GenBank_journal_distinct1에 업데이트 하였다. tb_GenBank_journal_distinct1의 titlesort와 tb_GenBank_journal의 titlesort를 조인하여 artctrlno를 tb_GenBank_journal 테이블에 업데이트 시킨 후 artctrlno의 건수를 산정하였다.

이렇게 논문 제목을 이용한 매핑 작업 수행 절차는 <그림 9>와 같다.

논문 제목의 가공을 통한 NDSL ID(문헌번호)의 연계를 위하여 NDSL과 GenBank 논문의 제목에서 특수문자와 공백을 제거하고, 제목을 소문자로 변환한 후, NDSL 논문 전체와 GenBank Pubmed id 미보유 논문에 대해서

동일 제목을 가진 NDSL ID를 추출한 결과 아래와 같은 결과를 얻었다.

- 중복을 제거한 논문 건수: 4,912만4,821건 => 6만8,515건
- 중복을 제거한 논문건수 중 NDSL 제목과 일치하는 건수: 1만9,370건
- 중복을 제거한 논문건수 중 제목과 발행연도가 일치하는 건수: 1만6,921건
- Pubmed id 중복 제거 건수: 17만5,989건

논문 제목을 통한 연계 시의 문제점은 제목이 동일한 다른 논문이 존재할 가능성이 있다는 것인데 이를 해결하기 위하여 논문 제목 이외에 저널 제목 및 발행연도를 추가로 비교하여 분석하였다. 분석을 위해 GenBank 참고 문헌 테이블의 저널 정보에서 연도를 추출하여 NDSL 저널 테이블의 연도와 1차 비교하였으며, NDSL 저널 발행연도와 GenBank 저널 정보의 발행연도를 비교하여 2차 비교 검증한 결과 전체 1만9,370건의 중복제거 논문 중 1

<그림 9> 논문 제목을 이용한 NDSL ID 매핑 방법

만6,921건의 데이터가 발행연도가 일치한다는 결과를 얻었다(〈그림 10〉 참조).

지금까지의 분석 결과, 논문 제목을 통한 데이터의 연계는 제목 간의 맵핑 성공률이 28.27%로 나왔지만(19,370/68,515*100) 저널 발행 연도를 추가로 비교하여 검토한 결과 전체 중복제거 논문건수를 기준으로 약 24.70%의 맵핑 성공률(16,921/68,515*100)을 기록하여 정확도는 87.36%(16,921/19,370*100)로 나타났다. 위의 작업을 통한 검토 결과 논문 제목을 통한 데이터의 연계는 상당히 높은 정확도(약87%)를 가지며 데이터 연계 시에 효과적으로 사용이 가능할 것으로 판단된다.

3.2.2 NDSL 특허 콘텐츠

GenBank reference의 journal 필드가 patent로 시작되는 데이터를 추출하고 가공하여 locus id, 특허명칭, 출원국가, 특허번호 등의 정보를 분리 저장하여 특허 테이블을 생성하였다. 〈표 4〉는 국가별 특허건수 현황이다. 건수로는 국제특허가 가장 많지만 중복을 제거하면 미국특허가 가장 많은 것으로 나타난다.

특허 데이터는 'Patent:'로 시작되며 국가코드+스페이스+특허번호+스페이스+공개일(등록일)+세미콜론(:)+출원인으로 구성되어 있다(예) Patent: US6,133,00817-OCT-2000; NewEnglandBiolabs,Inc.;Beverly,MA;US

〈그림 10〉 논문 맵핑 데이터의 검증

〈표 4〉 국가별 특허건수

국가코드	건수	비율	중복제거건수	비율
WO(WIPO)	2,084,558	46.28%	24,600	28.24%
US	1,262,903	28.04%	31,529	36.19%
JP	848,400	18.84%	25,169	28.89%
EP	304,433	6.76%	5,823	6.68%
기타	3,814	0.08%	-	
합계	4,504,108	100%	87,121	100%

※ 중복제거건수는 특허번호의 중복을 제거한 건수임.

A:). 특허 데이터의 국가코드, 특허번호, 공개일(등록일), 공개연도(등록연도)를 분리하여 별도의 컬럼에 저장한 후 NDSL 특허와 연계되도록 하였다. 특허번호의 길이에 따른 데이터 샘플을 10개 이상 분포된 데이터에 대해 KISTI의 NDSL 특허 데이터베이스와 비교한 결과 GenBank 참고문헌 데이터의 가공을 통해 데이터 맵핑이 가능하다는 것을 알 수 있었다.

〈표 5〉는 GenBank에서 추출한 450만 4,108 건의 특허 데이터에서 중복을 제거하고 특허번호 자리수별 현황을 분석한 결과이다. 〈표 5〉에서 보면 자리수의 대부분이 7자리, 8자리, 10자리에 분포되어 있다는 것을 알 수 있다. 이 분석 결과를 바탕으로 데이터 맵핑을 위하여 국가별 특허번호 자리 수에 따른 데이터 가공 방법을 정리한 것이 〈표 6〉이다.

〈표 5〉 특허번호 자리수별 현황

자리수	WO	US	JP	EP
1	1	1		
4		1		
5		7	2	
6				1
7	10,967	31,499		5,318
8	3,940	2	1	503
9		4		1
10	9,691		25,165	
11		3		
12	1	4		
13		6		
14		1		
15		1		
합계	24,601	31,529	25,169	5,823

〈표 6〉 국가별 특허번호에 따른 가공 방법

구분	자리수	맵핑 가공 방법	적용건수	비고
WO(WIPO)	7	현재 상태 링크 가능	803,261	공개번호
	8	첫 자리 숫자 제거	302,879	공개번호
	10	처음3자리 숫자제거	978,415	공개번호
	소계		2,084,555	
US	7	현재상태 링크가능	1,262,831	등록번호
JP	10	공개일자+공개번호 = JPA공개연월 + '0' + 특허번호 뒤에서 6자리 사용해서 조회 가능	848,400	NDSL관리 번호 검증 검색 시에는 뒤에서 7자리
EP	7	공개번호 = EP-앞에서 7자리	301,650	공개번호
	8	공개번호 = EP-앞에서 7자리	2,781	공개번호
	소계		304,431	
합계			4,500,217	

※ 일본특허의 경우 공개번호만으로 맵핑할 경우 1:n의 결과가 나오며, 이를 해결하기 위해서 NDSL 관리번호로 가공한 결과를 저장하였음.

특허 데이터 연계를 위해 <표 6>의 방법에 의해서 데이터를 가공하여 <표 7>의 테이블 (tb_GenBank_mapping)에 로딩하였다. 맵핑 건수는 450만217건이며, 전체 건수 대비 약 99.9%의 데이터가 맵핑 가능하였다. 맵핑 테이블의 flag 값에는 특허를 표시하는 P를 입력하면 된다.

3.2.3 KBIF 생물다양성 콘텐츠

GenBank와 생물다양성 콘텐츠(KBIF) 연계를 위해 사용될 테이블은 <그림 11>의 ERD에서 보는 것과 같이 KBIF 색인을 위한 테이블(indexing), 종정보 중복 제거 테이블(tb_gbif_taxon_distinct), GBIF taxonomy 원본 저장 테이블(tb_gbif_taxon_dump), GenBank 정보 저장 테이블(tb_GenBank),

GenBank organism 중복 제거 테이블(tb_GenBank_organism_distinct), scientific name 중복 제거 테이블(tb_indexing_distinct), GenBank와 KBIF 맵핑 테이블(tb_GenBank_kbif_mapping) 등 6개이다.

KBIF 데이터 저장소에는 13개 정보제공 기관에서 제공한 생물다양성 메타데이터를 통합하여 국가생물다양성정보포털(이하 NABIPOS)에서 서비스하고 있다. NABIPOS에는 13개 데이터 제공자(국가자연사연구종합정보시스템, 계룡산자연사박물관, 전남해양수산과학관, 산림청, 제주특별자치도 민속자연사박물관, 국가생물자원정보관리센터(KOBIC), 이화여자대학교 자연사박물관, 목포자연사박물관, 서대문자연사박물관, 경희대학교 자연사박물관, 국

<표 7> 맵핑 테이블(tb_GenBank_mapping) 구조

일련번호	컬럼명	데이터타입	크기	Null(Y/N)	색인
1	locus	varchar	15	N	Y
2	ndsl_id	varchar	30	N	Y
3	flag	char	1	N	Y

<그림 11> GenBank와 KBIF 데이터 연계 ERD

립수목원, 한남대자연사박물관, 한국과학기술 정보연구원)의 데이터를 통합하여 색인한 테이블을 별도로 관리하고 있다. NABIPOS의 색인 테이블은 검색에 필요한 정보와 원본 데이터의 위치를 포함하고 있으며 학명(scientific name)이 테이블에 저장되어 있다. 연계를 위하여 NABIPOS의 114만5,785건의 색인 테이블을 입수하여 Mysql에 업로드하였다.

이렇게 입수한 KBIF 데이터의 scientific name과 GenBank의 organism 데이터의 조인을 시도하였으나 조인한 결과 맵핑이 단 1건도 이루어지지 않았다. 이는 대소문자 구분, 학명 후반부의 발견자 정보, 구분정보 부여,

마침표(.) 등의 차이에 의한 오류로 인하여 조인이 이루어지지 않았던 것으로 밝혀졌다. 이를 해결하기 위하여 GBIF의 taxonomy 원본 데이터를 기준으로 KBIF의 scientific name과 GenBank의 organism 데이터를 정제하여 각 테이블에 저장하였다.

scientific name의 정제를 위하여 GBIF 데이터 375만1,737건을 <http://tools.gbif.org/>에서 다운로드 받아 Mysql에 업로드하였다. 이 중에서 중복을 제거하면 368만4,444건이 추출되며 중복 제거 데이터를 기준으로 맵핑을 위한 데이터베이스 구축 작업을 하였다. <표 8>은 GenBank와 KBIF의 맵핑 결과

<표 8> GenBank-KBIF 분석 결과

학명	한국명	건수	비율
zea mays	옥수수	2,878,332	15.73%
sus scrofa	멧돼지	2,121,442	11.59%
arabidopsis thaliana	애기장대	1,949,950	10.66%
oryza sativa	벼	1,587,998	8.68%
triticum aestivum	밀	1,104,170	6.03%
sorghum bicolor	단수수	1,011,769	5.53%
oryzias latipes	송사리	963,879	5.27%
glycine max	콩	766,948	4.19%
drosophila melanogaster	노랑초파리	740,537	4.05%
brassica napus	유채	582,964	3.19%
vitis vinifera	포도나무	533,617	2.92%
solanum tuberosum	감자	375,399	2.05%
pinus taeda	테다소나무	336,600	1.84%
gasterosteus aculeatus	큰가시고기	289,744	1.58%
bombyx mori	누에	279,610	1.53%
nicotiana tabacum	담배	179,282	0.98%
phaseolus vulgaris	강낭콩	118,651	0.65%
hordeum vulgare	보리	102,964	0.56%
helianthus annuus	해바라기	98,239	0.54%
citrus sinensis	오렌지	96,079	0.53%
artemisia annua 외 생물종		2,181,486	11.92%
합계		18,299,660	100.00%

를 생물 중별로 분석한 내용이다.

GBIF의 taxonomy 원본 데이터의 scientific name을 기준으로 KBIF의 scientific name을 데이터베이스에서 조회하고 실패했을 경우 scientific name을 마지막에서 1자리 제거 후 다시 데이터베이스에 조회하고 조회에 성공할 경우 필드에 성공한 scientific name을 저장하였다. GenBank 데이터의 건수가 약 8,400만 건이기 때문에 프로그램의 실행 시간을 최소화하기 위하여 GenBank의 organism 필드의 중복을 제거하여 별도의 테이블에 저장한 후 작업을 수행하였다. <그림 12>는 GenBank-KBIF 맵핑 절차를 나타낸 것이다.

<그림 12>와 같이 tb_GenBank_distinct_organism 테이블을 처리한 결과는 다음과 같다. 개발된 프로그램을 이용하여 tb_GenBank_distinct_organism 테이블의 organism 데이

터와 gbif_taxonomy 테이블의 scientific_name 데이터를 처리한 결과 전체 31만5,184건(organism의 중복제거 건수) 중 25만5,844건의 데이터 정제(약81.2%)에 성공하였다. 정제 완료된 데이터는 tb_GenBank 테이블에 별도의 필드(scientific_name)를 추가하여 업데이트하였다.

업데이트 시 전체 데이터 약 8,400만 건과 31만 건의 조인 업데이트는 서버 용량의 한계로 완료 시간을 예측할 수 없어 locus의 첫 자리를 구분하여 (A, B, C, ..., Z) 알파벳순으로 26번 처리하였는데 약 24시간이 소요되었다. tb_GenBank 테이블 업데이트 결과 전체 8,411만 2,248건의 데이터 중 8,077만 3,010건의 scientific_name이 정제되어 약 96%의 성공률을 보였다.

개발된 프로그램을 이용하여 tb_indexing_

<그림 12> GenBank-KBIF 맵핑 방법

distinct(KBIF_distinct) 테이블의 데이터를 처리한 결과 전체 2만6,727건 가운데 2만6,222건의 데이터가 정제되어 98.1%의 성공률을 보였다. 정제가 완료된 데이터는 KBIF(indexing) 테이블의 필드(scientific_name)에 업데이트하였다. 업데이트 결과 전체 114만5,785건 중 113만1,817건의 scientific_name이 업데이트되어 98.8%의 성공률을 보였다.

GenBank와 KBIF 데이터 정제가 완료된 후 GenBank organism 데이터와 KBIF의 scientific name의 조인결과 1,829만9,660건의 데이터가 조인되었으며 이는 전체 8,411만2,248건 중에서 21.76%가 조인되었다는 결과이다. 조인이 성공된 scientific name의 중복 제거 결과 4,698건이 나오며 이는 KBIF 중복제거 데이터 2만6,727건 대비 약 17.6%이다.

4. 콘텐츠 연계 프로토타입 구현

지금까지 GenBank 데이터 중에서 reference 필드를 분석하여 그 결과를 바탕으로 기존 KISTI 콘텐츠(논문정보, 특허정보, 생물다양성정보)와 GenBank를 상호 연계하기 위하여 매핑 기술을 이용하여 맵핑 테이블을 설계하여 KISTI CCBB 웹사이트(<http://www.cccb.re.kr>)에서 서비스 중인 GenBank 데이터베이스를 중심으로 연계 서비스 방안을 설계하였다. 본 장에서는 설계된 서비스 방안을 바탕으로 시스템간 연계 구조를 비롯하여 프로토타입을 구현하였다.

4.1 콘텐츠 연계 기법

4.1.1 연계 구조

본 서비스 시스템에서는 NDSL 논문, NDSL

〈그림 13〉 각 시스템간 연계 구조

〈그림 14〉 시스템간 연계를 위한 주요 키

특허, GenBank, KBIF 등 크게 4개의 서비스 시스템에서 제공하고 있는 콘텐츠 간의 연계를 통한 프로토타입 시스템을 개발하였다. 〈그림 13〉과 〈그림 14〉는 각 시스템간 연계 구조 및 연계를 가능하게 하는 주요 키를 나타내고 있다.

4.1.2 연계 기법

〈그림 15〉에서 보는 바와 같이 시스템 연계를 위하여 첫 번째로 NDSL 논문과 Genbank reference 필드의 title(논문명) 간의 맵

핑을 통한 NDSL ID 추출 및 연계를 수행하였다. 두 번째로 NDSL 특허와 GenBank의 reference 필드 중 patent 부분에서 추출한 특허번호를 정제하여 정제된 특허번호와 NDSL 특허번호를 연계하여 특허정보의 서지 및 원문정보를 검색 가능하게 하였다. 세 번째로 KBIF의 생물다양성정보에서 검색되는 학명(scientific name)과 GenBank source 필드 중 organism 부분의 중(taxonomy) 정보를 연계하였다.

〈그림 15〉 연계 서비스 프로세스 흐름도

4.2 콘텐츠 연계 인터페이스 구현

4.2.1 GenBank-NDSL 논문-NDSL 특허-KBIF

연계 인터페이스 구현

GenBank 데이터 필드는 locus(유전자 위치), definition(생물학적 특성), accession(고유 식별자), keywords(서열의 핵심어), source(생물체의 일반명/학명), reference(인용된 논문)

문사항), comment(주석 및 해설), features
(생물학적 특징정보), origin(서열의 원 소스)
등 크게 9개로 구성되어 있다.

〈그림 16〉과 같이 KISTI CCBB 웹사이트 GenBank 페이지에서 키워드(예)ginseng) 검색을 수행하면 〈그림 17〉과 같이 키워드와 관련된 간략보기 목록을 보여준다. 간략보기 목록 중 하나를 선택하면 유전자에 관한 상세 정

〈그림 16〉 GenBank 키워드 입력 화면

GenBank 검색시스템 :: 바이오인포매틱스 도구 :: Microsoft Internet Explorer

[홈](#)
[연락처](#)
[문의](#)
[FAQ](#)
[사이트맵](#)
[로그인](#)
[회원가입](#)
[도움말](#)

[http://dev.cdb.i.ri.or.kr/genbank/genbank_result.jsp?db=DEF&docID=genseng](#)

바이오인포매틱스센터
Center for Bioinformatics

GenBank 검색시스템

기분 DB	<input type="text" value="선택하여 기분 정보"/>	<input checked="" type="text" value="검색"/>	<input type="text" value="보통화환"/>
			총 2882건의 결과 중 1 - 20 번째 항목
순서	ACCESSION	제목 정보	
1	AB090929	Panax ginseng OSCPNK19 mRNA for Cycloartenol Synthase, complete cds. OSCPNK19; Cycloartenol Synthase. AB090929.1 GI:36486567	
2	AB011681	Synthetic oligonucleotide for differentiating varieties of medicinal carrot and base sequence. JF 2002.16806c A/L AB011681.1 GI:22091870	
3	DQ301876	2-5 Panax ginseng library (Lin W) Panax ginseng cDNA, mRNA sequence. EST. DQ301876.1 GI:83582419	
4	DQ301878	2-2 Panax ginseng library (Lin W) Panax ginseng cDNA, mRNA sequence. EST. DQ301878.1 GI:83582418	
5	DQ301874	2-4 Panax ginseng library (Lin W) Panax ginseng cDNA, mRNA sequence. EST. DQ301874.1 GI:83582417	
6	DQ301873	2-1 Panax ginseng library (Lin W) Panax ginseng cDNA, mRNA sequence. EST. DQ301873.1 GI:83582416	
7	DQ301872	1-10 Panax ginseng library (Lin W) Panax ginseng cDNA, mRNA sequence. EST. DQ301872.1 GI:83582415	
8	DN153790	Pv-c173 Phaseolus vulgaris (variety GAC Rex) cDNA Phaseolus vulgaris clone Pv-c173 '3' similar to Panax ginseng chloroplast, complete genome AY582139; similar to Phaseolus coccineus PCC61154 Scdript Runa Bean Suppressor Transf2 cDNA '5' similar to Chloroplast gene sequence CNA913777, mRNA sequence. EST. DN153790.1 GI:59873552	

〈그림 17〉 GenBank 간략보기 목록 화면

보를 보여 준다. 상세보기에 나타난 유전자와 관련되는 NDSL 논문, NDSL 특허, KBIF 생물다양성정보가 있을 경우 상세보기 상단에 관련 정보가 있는 시스템의 버튼이 나타나게 된다. 나타난 버튼 중에서 사용자가 원하는 정보의 버튼을 클릭하면 원하는 콘텐츠가 있는 시스템으로 연계되도록 인터페이스를 구현하였다.

콘텐츠간 연계를 위한 맵핑 테이블에 맵핑 데이터베이스를 구축한 결과 NDSL 논문, NDSL 특허, KBIF 생물다양성정보 3개의 콘텐츠가 모두 나타나는 경우는 1건도 없었다. 그러나 NDSL 논문과 KBIF 생물다양성정보, NDSL 특허와 KBIF 생물다양성정보가 동시에 나타나는 경우에는 <그림 18>, <그림 19>와 같이 연계되는 콘텐츠를 검색 가능하도록

<그림 18> GenBank-NDSL논문-KBIF 연계

<그림 19> GenBank-NDSL특허-KBIF 연계

인터페이스를 구현하였다. <그림 18>은 특정 필드를 추출하여 맵핑 테이블을 작성한 후 GenBank에서 검색한 유전자 정보에 NDSL 논문과 KBIF 생물다양성정보를 연계하여 서비스하는 화면이고, <그림 19>는 특정 필드를 추출하여 맵핑 테이블을 작성한 후 GenBank에서 검색한 유전자 정보에 NDSL 특허와 KBIF 생물다양성정보를 연계하여 서비스하는 화면이다.

4.2.2 NDSL 논문-GenBank 연계 인터페이스 구현

<그림 20>과 같이 NDSL 논문 웹사이트 (<http://scholar.kisti.re.kr>)에서 키워드 검색을 수행하면 <그림 21>에서 보는 바와 같이 키워드와 관련된 간략보기 목록을 보여준다. 간략보기 목록 중 하나를 선택하면 키워드와 관련된 논문에 관한 상세 정보를 <그림 22>와 같이 보여 준다. 상세보기에 나타난 논문과 관

<그림 20> NDSL 논문 키워드 입력 화면

<그림 21> NDSL 논문 간략보기 화면

〈그림 22〉 NDSL 논문에서 GenBank 데이터 보기

련되는 GenBank 유전자정보가 있을 경우에 검색결과 상단에 GenBank 데이터 보기 버튼이 나타난다. 나타난 버튼을 클릭하면 관련 유전자 간략보기 목록이 나타나고 목록 중에서 이용자가 원하는 정보의 버튼을 클릭하면 원하는 유전자 정보를 볼 수 있는 상세보기 화면으로 연계되도록 인터페이스를 구현하였다.

4.2.3 NDSL 특허-GenBank 연계 인터페이스 구현

〈그림 23〉과 같이 NDSL 특허 웹사이트 (<http://patent.kisti.re.kr>)에서 키워드 검색을 수행하면 〈그림 24〉에서 보는 바와 같이 키워드와 관련된 간략보기 목록을 보여준다. 간략보기 목록 중 하나를 선택하면 키워드와

〈그림 23〉 NDSL 특허 키워드 입력 화면

〈그림 24〉 NDSL 특허 간략보기 화면

〈그림 25〉 NDSL 특허에서 GenBank 데이터 보기

관련된 특허에 관한 상세 정보를 〈그림 25〉와 같이 보여 준다. 상세보기에 나타난 특허와 관련된 GenBank 유전자정보가 있을 경우에 검색결과 상단에 GenBank 데이터 보기 버튼이 나타난다. 버튼을 클릭하면 관련 유전자 간략보기 목록이 나타나고 목록 중에서 이용자가 원하는 정보의 버튼을 클릭하면 원하는 유전자 정보를 볼 수 있는 상세보기 화면으로 연결되도록 인터페이스를 구현하였다.

5. 결론

지금까지 GenBank 데이터베이스를 활용하여 각 필드를 분석하여 그 결과를 산출해 보았으며, 산출된 결과를 기본으로 GenBank와 NDSL 논문, NDSL 특허, KBIF 연계를 위한 맵핑 테이블을 설계하고 콘텐츠간 연계를 위한 메타 데이터베이스를 구축하였다. GenBank와 학술논문정보의 연계를 위해 Pub-

med id와 제목, 발행연도를 이용한 데이터 맵핑을 시도하였고, GenBank와 특허정보의 연계를 위하여 GenBank의 필드를 가공하여 특허번호를 추출하였다. GenBank와 생물다양성정보의 연계를 위해 GenBank의 organism 필드를 분석한 후 KISTI에서 서비스하고 있는 약 110만 건의 생물다양성 데이터의 종정보와 연계 가능하도록 맵핑 테이블을 설계하고, 설계된 내용을 바탕으로 프로토타입 시스템을 완성하였다.

본 논문에서 설계하여 구현한 프로토타입이 본 시스템으로 개발되어 KISTI NDSL에 적용되어 서비스 된다면 다음과 같은 기존 정보서비스와는 차별화된 다음과 같은 기대효과를 가져 올 것으로 사료된다. 첫째, GenBank에서 제공되지 않는 pubmed id 미보유 논문 중 상당수의 논문이 NDSL 학술논문정보 연계로 제공 가능할 것이다. 둘째, 미국 특허를 위주로 서비스되고 있는 GenBank와 KISTI에서 보유하고 있는 유럽 및 일본 특허 정보와의 연계가 가능하게 되어 유전정보 관련 특허정보가 폭넓게 검색될 것이다. 셋째, 국내 생물다양성정보인 KBIF와의 연계로 인하여 유전정보와 생물 종정보 검색이 한 번에 이루어져 생명과학 연구자들에게 더욱 유용한 고부가가치 서비스를 제공할 수 있을 것이다. 넷째, 이중 DB 간의 연계융합을 통한 고품질 학술정보 DB를 생성하여 향후 타 콘텐츠나 서비스와의 결합을 통한 매쉬업 서비스에 활용 가능할 것이다. 마지막으로 분야별 전문 이용자를 위한

주제기반 웹 포털로의 발전을 통한 주제 전문 정보서비스의 기반을 마련할 수 있을 것으로 기대된다.

그렇지만 이중의 콘텐츠를 자동으로 연계하는데 있어 논문 제목만을 가지고 맵핑하려다 보니 누락 또는 중복되는 데이터가 발생하였다. 논문의 발행년도를 추가로 적용하여 오류를 수정하였으나 기계적인 처리의 한계로 인하여 여전히 중복 또는 누락되는 데이터가 발생하였다. 또한 데이터 건수가 워낙 많다보니 중복을 제거하는 알고리즘이나 서버의 성능에 따라 작업시간에 많은 차이가 나는 것을 알 수 있었다. 현재 구현된 알고리즘으로는 Genbank에서 필요한 필드만 추려내는 데도 3일 이상의 시간이 걸렸다. 작업시간을 줄이는 데에도 많은 노력이 필요하리라 본다.

참고 문헌

- 김영기, 정종근, 이수상. 2007. 주제기반 온라인 학술 커뮤니티의 구축 방향. 『정보관리학회지』, 24(4): 5-31.
- 안부영, 오충식. 2008. 『생명정보 콘텐츠 업데이트 가이드 v. 2.0』. 대전: 한국과학기술정보연구원.
- 안부영, 한정민, 한진, 이상호. 2008. 생명정보 연계검색 인터페이스 설계에 관한 연구. 『제29회 한국정보처리학회 춘계학술발표대회 논문집』, 15(1): 407-409.
- 이상기, 최희운, 김선태, 이태석, 한희준, 현미환,

- 예용희. 2008. 이중 학술콘텐츠간 연계융합 사례연구-KISTI CLICK 중심. 『한국비블리아』, 19(1): 5-17.
- 이수정, 용환승. 2004. 웹서비스 기반 바이오서열정보 데이터베이스 및 통합검색시스템 개발. 『한국정보처리학회논문지』, 11D(4): 755-764.
- 과학기술정보 통합서비스 웹사이트. [인용 2009. 09.01].
〈<http://www.ndsl.kr/>〉.
- 국가생물다양성정보포털 웹사이트. [인용 2009. 09.01].
〈<http://nabipos.kisti.re.kr/>〉.
- 무기결정구조 DB(ICSD DB) 웹사이트. [인용 2009.09.01].
〈<http://icsd.kisti.re.kr/>〉.
- 플라즈마물성 DB(PLASMA DB) 웹사이트. [인용 2009.09.01].
〈<http://plasma.kisti.re.kr/>〉.
- 특허청 전통지식포털 웹사이트. [인용 2009. 09.01].
〈<http://www.koreantk.com>〉.
- 화합물 DB(CHEM DB) 웹사이트. [인용 2009. 09.01].
〈<http://chemdb.kisti.re.kr/>〉.
- 한국생물다양성정보기구 웹사이트. [인용 2009. 09.01].
〈<http://www.kbif.re.kr/>〉.
- KISTI 바이오인포매틱스 웹사이트. [인용 2009. 09.01].
〈<http://www.cccb.re.kr/>〉.
- KISTI 특허정보 웹사이트. [인용 2009.09.01].
〈<http://patent.ndsl.kr/>〉.
- KISTI 학술정보 웹사이트. [인용 2009.09.01].
〈<http://scholar.ndsl.kr/>〉.
- NCBI 웹사이트. [인용 2009.09.01].
〈<http://www.ncbi.nlm.nih.gov/>〉.