한글대장경 웹 검색 시스템의 구현

구현우*, 선수림*, 박미화*, 이재수**,
이용규***, 이금석***, 홍영식***, 한보광****

목 차

요 약

본 연구는 한글대장경 전산화 6차 사업으로 한글대장경 30책 분량을 전산화하여 검색 시스템을 구축하는데 목적이 있다. 고려대장경의 우리말 번역본인 한글대장경을 전산화하기 위해 개역된 고문헌을 입력하여 데이터베이스로 구축하고, 인터넷을 통해

*동국대학교 컴퓨터공학과
**동국대학교 불교학과, 전자불전・문화재콘텐츠연구소 전임연구원
***동국대학교 컴퓨터공학과 교수
****동국대학교 선학과 교수, 전자불전・문화재콘텐츠연구소 소장

그 내용을 검색할 수 있도록 한다. 한글대장경 고문헌은 확장한자, 누락문자, 특수문자 등을 포함하고 있어서, 본 연구에서는 효과적인 입력과 저장을 위해 유니코드(Unicode)를 사용하며, 유니코드로 표현하지 못하는 문자들은 이미지 폰트를 생성하여 표현한다. 데이터베이스를 구축하기 위해서 DBMS로는 MS-SQL 7.0을 사용하고, 운영체제로는 윈도우 2000 서버를, 웹 서버로는 IIS(Internet Information Server)를 사용하여 검색 시스템을 구축하였다. 또한 다양한 검색 방법을 제공하는 검색 엔진을 개발하여, 유니코드로 저장된 한글대장경 고문헌의 내용을 웹(http://ebtc.dongguk.ac.kr/)을 통해 보다 쉽게 전 세계에서 접근할 수 있도록 한다.

1. 서 론

본 연구는 한글대장경 전산화 6차 사업으로 한글대장경 30책, 110경을 전산화하여 전 세계에서 활발하게 사용되고 있는 인터넷을 통하여 검색할 수 있도록 하는 것이다.

불법이 인도에서 꽃을 피워 인류의 정신문화를 꽃피워 왔다. 불교의 가르침은 보통 사람들이 구사하는 언어를 통해 전해져왔는데, 초기에는 부처님으로부터 신성한 가르침을 직접 듣는 것이 가능하였고, 입에서 입으로 구전되어 왔다.

부처님의 입멸 후 그러한 가르침의 전통은 인도에서 결집(結集)을 통해 문자화되어 보다 많은 인류를 깨달음의 길로 이끄는 지침이 되었다. 부처님의 가르침은 동아시아의 거의 모든 국가에 전해졌고, 시대와 역사를 초월하여 각국의 찬란한 정신문화를 이끌어 왔다.

불교사는 경전의 역경을 통해 이루어져왔다고 해도 과언이 아니다. 경전의 번역과 수집 및 출간은 국가의 가장 큰 문화사업으로 중요시되어 왔다.

세계문화유산으로 등록된 고려대장경은 몽고의 침입으로 국가가 위기에 처했던 시기에 부처님의 가르침으로 국가의 안녕과 백성의 평안을 기원하기 위해 전 국가적으로 역량을 결집한 우리의 문화유산인 것이다.

조선시대에 이르러서는 훈민정음의 창제로 일반 백성들도 우리나라 말과 글을 널리 사용할 수 있게 되었다. 한문불경을 훈민정음으로 번역해 민간에 널리 유포시키기 위하여 간경도감에서 한글로 된 불경이 제작되기 시작하였다. 이는 지식인만의 불교에서 일체중생을 위한 불교로의 전환을 의미하게 된다. 조선 말기에서부터 가속화된 불경의 한글화는 일제의 강점기에 민족의 정신을 일깨우는 작업으로 진행되어 오늘에 이르게 되었다.

동국대학교의 역경원 설립과 함께 본격화되기 시작한 한글대장경 사업은 현대문명의 발달에 발맞추어 새롭게 전산화의 길을 모색하고 있다. 이는 한글대장경을 디지털화하여 인터넷을 통해 전세계의 인류에게 제공함으로써 시간과 장소를 초월하여 불법의 진리를 홍보하는 것이며, 또한 우리나라의 뛰어난 정신문화를 전세계에 알리는 새로운 전법활동이라고 할 수 있다.

2. 한글대장경 전산화 6차 사업

2.1 한글대장경의 입력 · 교정 · 색인작업

동국역경원에서 각 분야의 전문적인 능력을 지닌 역경위원들의 엄정한 번역을 담당하고, 그 번역된 결과를 입력하였다. 이에 대해 교정과 윤문을 거쳤고, 3차에 걸쳐서 엄밀한 교정작업을 수행하였다.

한글대장경 전산화 제 6차 사업은 2006년 6월부터 2007년 5월까지 수행하였다. 이번 사업에서 입력교정한 대장경의 목록은 총 30책 분량, 110경으로 다음과 같다. (※ K번호는 고려대장경의 경전고유번호임.)

K.0004 광찬경(1-10권)
K.0008 승천왕반야바라밀경
K.0010 문수사리소설마하반야바라밀경
K.0011 문수사리소설반야바라밀경
K.0012 불설유수보살무상청정분위경
K.0013 금강반야바라밀경
K.0015 금강반야바라밀경

전자불전 제9집(2007)

K.0016 능단금강반야바라밀다경(현장)
K.0017 능단금강반야바라밀다경(의정)
K.0019 인왕반야바라밀경
K.0021 마하반야바라밀대명주경
K.0024 무량청정평등각경
K.0025 아미타삼야삼불살루불단과도인도경
K.0029 불설보문품경
K.0034 불설환사인현경
K.0035 결정비니경
K.0036 수마제경
K.0037 발각정심경
K.0039 불설수마제보살경
K.0040 불설아사세왕녀아술달보살경
K.0041 불설이구시녀경
K.0042 득무구녀경
K.0043 문수사리소설부사의불경계경
K.0046 태자쇄호경
K.0047 태자화휴경
K.0048 혜상보살문대선권경
K.0049 대승현식경
K.0050 대승방등요혜경
K.0051 미륵보살소문본원경
K.0053 불설마하연보엄경
K.0054 승만사자후일승대방편방광경
K.0055 비야사문경
K.0081 신력입인법문경
K.0082 불화엄입여래덕지부사의경계경
K.0084 대방광불화엄경수자분
K.0085 도제불경계지광엄경
K.0086 대방광입여래지덕부사의경
K.0087 대방광여래부사의경계경
K.0088 대방광불화엄경부사의불경계분
K.0091 대방광보현소설경
K.0096 대방광보살십지경
K.0097 불설보살십주경
K.0099 여래흥현경
K.0102 불설라마가경
K.0104 대방광불화엄경입법계품
K.0105 대반열반경
K.0111 방광대장엄경
K.0113 불설법화삼매경
K.0114 무량의경덕행품
K.0115 살담분타리경
K.0135 불설아유월치차경

K.0136 광박엄정불퇴전륜경
K.0137 불퇴전법륜경
K.0147 제제방등학경
K.0148 대승방광총지경
K.0163 대살차니건자소설경
K.0188 여래장엄지혜광명입일체불경계경
K.0189 도일체제불경계지엄경
K.0240 불설보적삼매문수사리보살문법신경
K.0362 불설수뢰경
K.0369 불설보살수행경
K.0380 관보현보살행법경
K.0382 최승문보살십주제구단결경
K.0416 대법고경
K.0664 불설중본기경
K.0777 과거현재인과경
K.0800 불설우전왕경
K.0804 불설흥기행경
K.0829 불오백제자자설본기경
K.0957 아비달마장현종론(1-11권)
K.1009 불설십이유경
K.1029 문수사리발원경
K.1081 광홍명집(1-15권)
K.1100 불설대승일자왕소문경
K.1123 대가섭문대보적정법경
K.1139 관상불모반야바라밀다보살경
K.1152 묘비보살소문경
K.1172 중허마하제경
K.1185 불설비사문천왕경
K.1186 불설성관자재보살범찬
K.1189 불설변조반야바라밀경
K.1200 불설불모보덕장반야바라밀경
K.1206 불설호국존자소문대승경
K.1263 신화엄경론
K.1267 보변지장반야바라밀다심경
K.1275 대락금강불공진실삼마야경
K.1303 불설삼십오불명예참문
K.1340 인왕호국반야바라밀다경
K.1383 반야바라밀다심경
K.1414 성관자재보살공덕찬
K.1415 불설요의반야바라밀다경
K.1419 불설최승묘길상근본지최상비밀일체명의삼마지분
K.1423 불설불모출생삼법장반야바라밀다경(1-25권)
K.1424 대방광선교방편경
K.1434 불설무이평등최상유가대교왕경

K.1436 불설불모반야바라밀다대명관상의궤
K.1439 불설보대다라니경
K.1440 금신다라니경
K.1442 금강장장엄반야바라밀다교중일분
K.1443 불길상덕찬
K.1450 불설여환삼마지무량인법문경
K.1452 일체비밀최상명의대교왕의궤
K.1455 성팔천송반야바라밀다일백팔명진실원의다라니경
K.1461 광대발원송
K.1462 불설비밀상경
K.1468 불설무외수소문대승경
K.1485 개각자성반야바라밀다경
K.1487 대승보살장정법경
K.1489 대승입제불경계지광명장엄경
K.1491 십불선업도경
K.1493 사사법오십송

2.1.2. 태그 작업

입력과 교정을 마친 30책의 한글대장경은 본 연구소 입력팀에서 페이지·대제목·소제목·해제·서론·각주·진언이미지에 대하여 각각 태깅 작업을 수행하였다.

1) 페이지 태그작업 : 페이지를 검색하여 해당 원문을 보여주었다.
   페이지의 태그는 '<PAGE PAGENUM='경번호-10001'/>'로 하였다.
2) 제목 태그작업 : 경전의 대제목과 소제목을 검색할 수 있으며, 경전의 제목을 통하여 해당 원문을 확인할 수 있게 하였다.
   경 제목의 태그는 '<JMOK1 SEARCH='TRUE'>경 제목</JMOK1>'로 하였다.
   경의 소제목의 경우 태그는 '<JMOK2>경 소제목</JMOK2>'과 그 이하의 소제목은 '<JMOK3>경 소제목</JMOK3>'로 하였다.
3) 각주 태그작업 : 한글대장경의 각주에 나타나 있는 원문을 확인할 수 있도록 하였다.
   각주 태그는 원문의 각주는 '<COM NUM='1'/>'로 하며, 각 해당 페이지의 하단에
   '<COMMENT> 각주의 내용</COMMENT>'으로 입력하였다.

4) 진언 이미지 태그작업 : 한글대장경에 나타난 진언은 이미지로 처리한 후 이미지로 확인할 수 있도록 하였다.
진언 이미지의 태그는
'<JIN 경 번호-페이지-페이지의 진언 일련번호>
진언내용 </JIN>'로 하였다.

2.2 데이터베이스 저장

한글대장경을 데이터베이스로 저장하고 관리하기 위해서는 원문을 구별해 주는 각 태그들의 유효성을 검증하는 작업과 원문으로부터 제목, 원문 내용, 키워드를 추출하여 유니코드로 변환하고 해당 테이블에 값을 저장하는 작업이 필요하다. 또한 데이터베이스 저장과 관리를 위한 저장 시스템이 요구된다. 본 절에서는 제6차 사업에서 수행한 데이터베이스 관련 작업과 한글대장경 데이터베이스 구축 단계, 데이터베이스 저장 시스템의 구성과 기능, 데이터베이스 구조에 대해 기술한다.

2.2.1 데이터베이스 부분 개선 내용

한글대장경 제6차 사업에서는 새로운 한글대장경을 데이터베이스로 구축하였으며 기존에 구축한 데이터베이스에 대한 재검증 작업을 진행하였다. 또한 데이터베이스 검색 속도 개선을 위하여 웹 검색에 사용된 데이터베이스 검색 질의를 재검증 및 수정하였다. 이를 좀 더 자세히 기술하면 다음과 같다.

첫째, 새로운 한글대장경 110경을 데이터베이스로 구축하였다. 현재 한글대장경 사업을 통해 구축된 데이터베이스에는 한글대장경 464경이 포함되어 있다.

둘째, 기존에 구축한 한글대장경 원문을 대상으로 전체적인 재검증 작업을 수행하였다. 재검증 작업을 통하여 오류가 있었던 88경의 한글대장경 원

문을 수정하였다. 기존 경의 재검증 작업은 다음과 같이 수행되었다.

① 누락된 고려·신수번호의 보완 작업
② 특수 기호의 깨짐 현상 보정
③ 잘못된 경명으로 인한 검색 오류 보완 작업
④ 중복된 경과 누락된 페이지 정정 작업
⑤ 누락 및 중복 페이지 내용 수정

셋째, 데이터베이스 검색 속도 개선을 위하여 웹 검색에 사용된 데이터베이스 검색 질의를 재검증 및 수정하였다. 이를 통해 키워드 검색에서의 서비스 시간을 단축시킬 수 있었다.

넷째, 전체 데이터베이스를 검색하여 고려번호와 신수번호가 없는 경들을 모두 조사한 후 입력 팀과 협력하여 모든 경들의 고려번호와 신수번호를 재검증하고 수정 및 입력하였다.

2.2.2 한글대장경 데이터베이스 구축 단계

한글대장경의 원문은 제목, 원문, 주석 등으로 구성되어 있고, 각각에 해당되는 내용은 태그로 구별하여 데이터베이스를 구축한다. 원문에 나타나는 한자는 기존 문자 집합으로 표현하는데 한계가 있어 유니코드로 변환하여 저장한다. 한글대장경 데이터베이스를 구축하기 위하여 먼저, 원문을 구별해 주는 각 태그들의 유효성을 검증하는 작업을 진행한다. 유효성 검증 작업 후 원문으로부터 제목, 원문 내용, 키워드를 추출하여 유니코드로 변환한 후 그 값을 해당 테이블에 저장한다.

(1) 태그가 삽입된 원문의 유효성 검증 작업

텍스트 파일로 변환된 원문에는 제목, 페이지, 주석, 진언 이미지 등을 구별하기 위하여 각각 〈JMOK〉, 〈PAGE〉, 〈COMMENT〉, 〈IMG〉라는 태그들을 삽입한다. 이러한 태그들은 여는 태그(〈...〉)와 닫는 태그(〈/...〉)가 쌍으로 구성되어야한다. 태그가 잘못 작성되면 잘못된 데이터가 데이터베이스에 입력될 수 있으므로 유효성 검증 작업이 필요하다. 원문의 유효성 검증

작업은 다음과 같은 순서로 이루어진다.

① "*.txt"로 저장된 원문 파일을 "*.xml"로 확장자명을 변경한다.

② 웹 브라우저에서 해당 XML 문서를 불러들인다.

③ 웹 브라우저에 에러 메시지가 나타나지 않으면 유효한 문서이고, 에러 메시지가 나타나면 해당되는 내용을 찾아 원문을 수정한다.

(2) 키워드 인덱스 생성 및 저장

키워드 추출 및 저장 단계에서는 원문 내에서 키워드로 지정된 단어를 찾아 그 위치와 단어를 keyword_index 테이블에 저장한다. 미리 지정된 키워드 목록 정보는 keyword 테이블에 유니코드로 저장되어 있다.

유니코드를 검색하기 위해서는 기존의 아스키파일과는 달리 2바이트 단위로 문자를 비교해야 한다. 키워드 인덱스 구축에는 원문이 저장된 "edocdata"와 미리 지정된 키워드 정보를 갖는 "keyword" 테이블이 사용된다. 키워드 인덱스를 구축하는 절차는 다음과 같다.

① 키워드 테이블에서 하나의 키워드를 추출하고, "edocdata" 테이블의 전체 내용과 비교한다.

② "keyword" 테이블에서 선택한 키워드와 같은 코드를 발견하면, 일련번호(uid), 키워드번호(keynum), 키워드(keyword), "edocdata" 테이블에서의 페이지(pagenum) 정보를 keyword_index 테이블에 저장한다.

③ "edocdata" 테이블의 모든 레코드를 비교한 다음에 다음 키워드를 읽어 키워드 비교 과정을 반복 수행한다.

④ "keyword" 테이블 전체를 비교했을 때, 프로그램을 종료한다.

키워드 저장은 키워드가 저장된 텍스트 파일로부터 키워드를 추출하여 키워드 테이블을 구축하는 방법을 이용한다. 키워드 파일에는 한글 키워드와 한자 키워드가 저장되어 있으며, 실제 "keyword" 테이블에는 한글과 한자 키워드에 대한 유니코드 값이 저장된다.

(3) 제목 인덱스 구축

원문에 〈JMOK〉과 〈/JMOK〉 태그가 나타나면 해당 제목의 레벨과 위치를 "tag_jmok_area_table" 테이블에 저장한다.

제목 태그는 트리 구조의 형태를 갖는다. 〈JMOK〉 다음에 나타나는 숫자인 1, 2, 3, 4는 제목의 레벨을 나타내고, 〈JMOK4〉 태그의 속성인 "SERCH='TRUE'"는 경 제목을 의미하며 그 값은 "tag_kyung_table"에 저장한다.

(4) 원문 저장

원문 저장은 유니코드 편집기에서 작성된 유니코드 원문을 줄 단위로 읽어 "edocdata" 테이블에 저장한다. "edocdata" 테이블에는 페이지 당 라인(line)수와 단 번호 등의 부가 정보를 저장한다. 부가 정보는 원문에 대한 인덱스 역할을 한다. 원문을 저장할 때 원문에서 한 라인이 레코드의 저장 크기를 초과할 경우에 100자 단위로 나눠서 저장하게 되는데 이때 라인이 이어짐을 "ncontinue" 속성으로 표현한다.

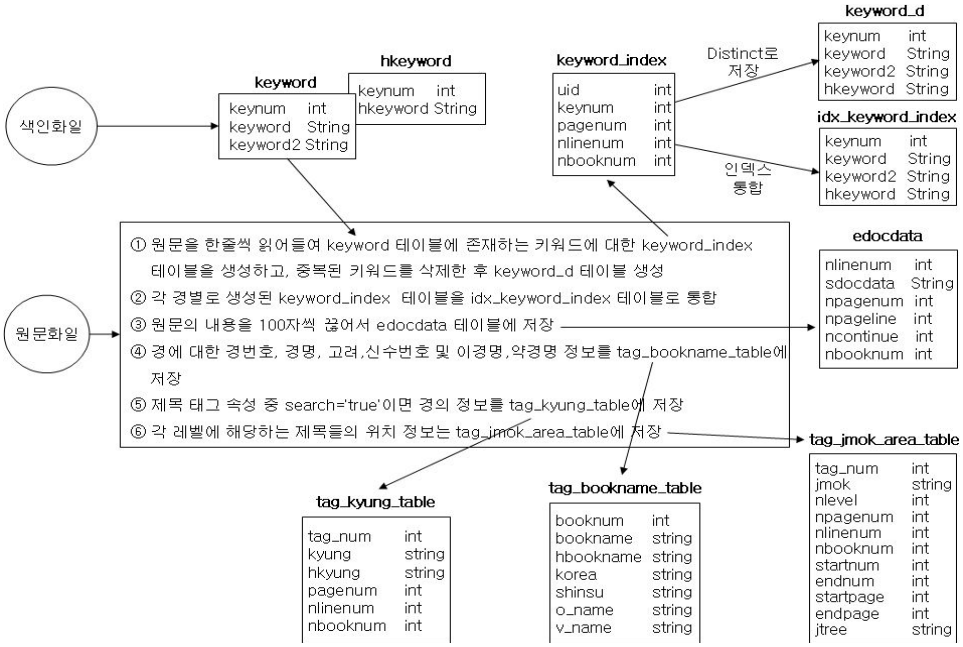
2.2.3 데이터베이스 구축

데이터베이스 구축을 위해 마이크로소프트 SQL Server 7.0 DBMS를 사용하였다. [그림 1]은 색인 파일과 원문 파일을 이용하여 각각의 테이블을 생성하는 방법과 테이블 구성정보를 보이고 있다.

테이블 생성 과정을 요약하면 다음과 같다.

- 색인 파일을 읽어서 "keyword"와 "hkeyword" 테이블을 만든 다음 원문에서 읽어 들인 한 줄에 keyword 테이블의 키워드가 존재하면 그 키워드는 "keyword_index" 테이블에 저장한다.
- 원문은 100자씩 끊어서 "edocdata" 테이블에 저장한다.
- 고려·신수번호와 이경명·약경명 정보는 "tag_bookname_table"에 저장한다.

● 제목 태그 중 "SERCH='TRUE'" 인 제목은 "tag_kyung_table" 테이블에 저장한다. 각 제목 태그 정보는 각 레벨 정보와 위치 정보를 "tag_jmok_area_table"에 저장한다.



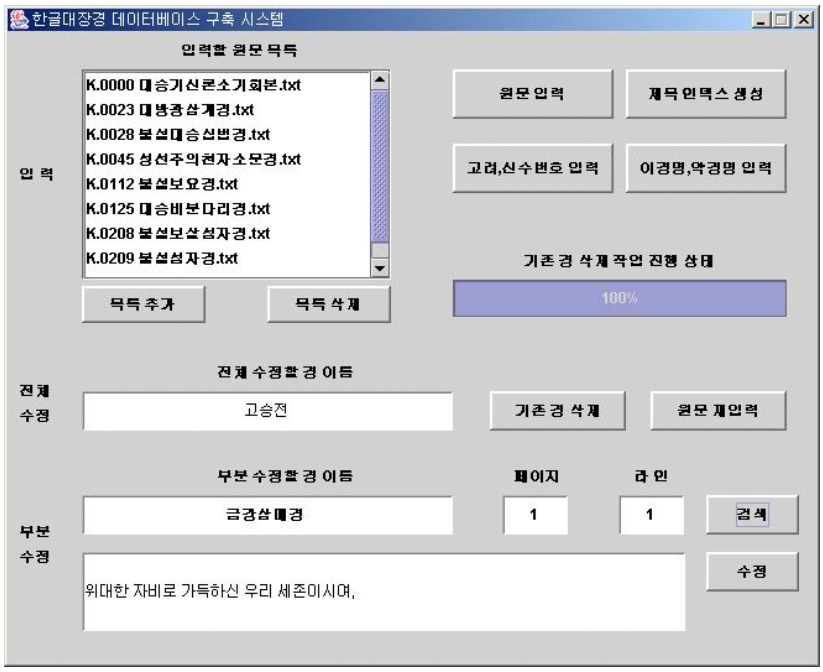
[그림 1] 테이블 생성 방법

2.2.4 데이터베이스 저장 및 관리 시스템

데이터베이스 저장 및 관리 시스템은 한글대장경 원문의 내용을 입력·수정·삭제하는 기능을 지원하며, 원문을 부분적으로 검색하여 수정할 수 있도록 한다. [그림 2]는 데이터베이스 저장 및 관리 시스템의 주화면이다.

원문을 입력하기 위해서는 왼쪽 중간에 있는 '목록 추가' 버튼을 눌러 '입력할 원문 목록'에 추가한 뒤, '원문 입력' 버튼을 누른다. 원문이 저장되는 과정은 오른쪽 중간에 있는 '원문 입력 작업 진행 상태' 그래프를 통하여 확인할 수 있다. 오른쪽 상단에 있는 '제목 인덱스 생성' 버튼을 누르면 각 레벨에 해당하는 제목들의 위치 정보를 갖는 제목 인덱스가 생성된다. '고려, 신수번호 입력' 및 '이경명, 약경명 입력' 버튼을 누르면 생성된 제목 인덱스 테이블에 해당 정보들이 입력된다. 원문 수정을 위해서는 전체 수정할 경 이름을 입력한 뒤, 오른쪽 중간에 있는 '기존 경 삭제' 버튼을 눌러 기존의

원문 전체를 삭제한 뒤, '원문 재입력' 버튼을 눌러 기존 경의 전체 수정할 원문을 재입력한다.



[그림 2] 한글대장경 데이터베이스 저장 및 관리 시스템

데이터베이스 저장 및 관리 시스템에는 또한 저장된 내용을 검색하거나 수정할 수 있는 기능이 제공된다. 검색 및 수정하고자 하는 경 이름, 페이지, 라인 정보를 입력한 뒤에 오른쪽 하단에 있는 '검색' 버튼을 누르면 해당 원문이 검색되고, 수정할 내용을 입력하여 '수정' 버튼을 누르면 부분 수정이 이루어진다.

2.3 웹 검색 인터페이스

한글대장경 웹 검색 인터페이스는 사용자가 웹을 통하여 한글대장경을 보다 편리하게 검색할 수 있도록 다양한 검색 방법을 제공하고 있다. 검색 방법은 크게 경명검색, 용어검색, 쪽수검색으로 구성되어있으며 각각의 검색

방법으로 검색한 결과가 한글대장경의 어느 부분에 속하는지를 쉽게 알 수 있도록 위치 정보를 제공하고 있다. 또한 사용자가 한글대장경을 검색한 후 그 페이지에 해당하는 고려대장경과 신수대장경의 권수와 페이지 정보를 제공하고 있다.

또한 사용자 접근성을 높이기 위한 검색 사용자 인터페이스 및 기능을 지속적으로 개선하였다. 개선된 검색 사용자 인터페이스 및 기능은 첫째로 용어검색에서 분리되어 있던 기본검색과 상세검색 인터페이스의 사용성을 높이기 위하여 통합된 인터페이스로 개선하였으며, 둘째로 검색된 본문 화면에 프레임을 적용하여 본문 내용을 편리하게 이용할 수 있는 기능들을 상단에 배치하였고, 셋째로 화면 확대/축소 기능을 추가하여 사용자가 원하는 대로 화면 크기를 조정할 수 있도록 함으로써 내용의 가독성을 높였으며, 마지막으로 검색된 본문 페이지 내의 내용을 검색할 수 있는 기능을 추가하여 원하는 정보를 빠르게 찾을 수 있도록 하였다.

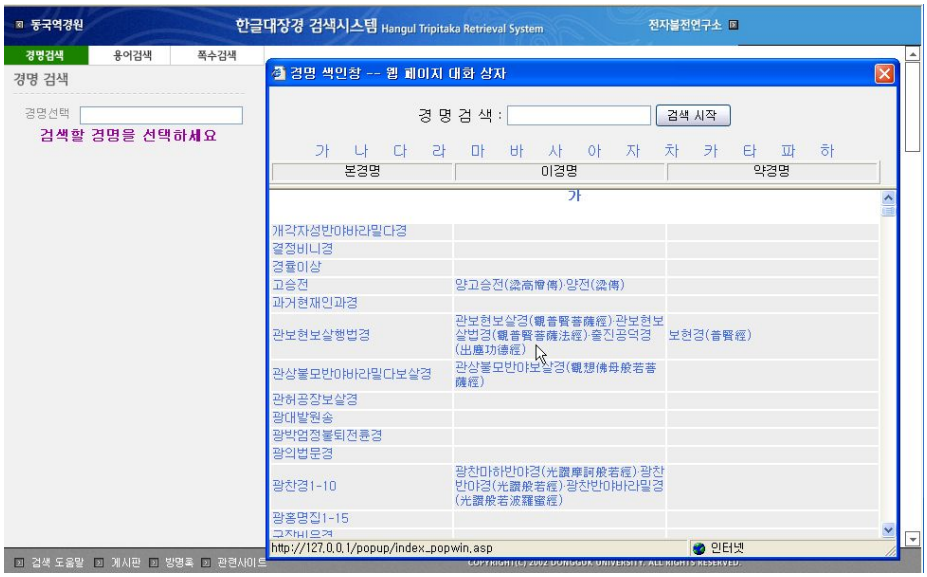
웹 검색 인터페이스의 주요 기능에 대한 자세한 내용은 본문을 통하여 살펴보도록 한다.

2.3.1 웹 검색 인터페이스 주요 기능

한글대장경 웹 검색인터페이스는 사용자가 편리하게 검색 할 수 있도록 여러 가지 검색 방법을 제공하고 있다. 주요 검색 방법은 경명검색, 용어검색, 그리고 경명쪽수검색 등이 있다. 보다 자세한 내용은 다음과 같다.

(1) 경명검색

경명검색은 〔그림 3〕과 같이 경명 목록에서 특정 경명을 선택하여 그 경에 대한 내용을 검색하는 방법이다. 한글대장경의 많은 경전들은 본경명(本經名) 뿐만 아니라 이경명(異經名)과 약경명(略經名)도 가지고 있다. 본경명은 대장경을 잘 아는 전문가 위주의 경명이고, 이경명이나 약경명은 일반 사용자에게 친숙한 경명이다. 본 검색시스템은 본경명은 물론 이경명이나 약경명으로도 쉽게 검색 할 수 있는 서비스를 제공하고 있다. 이와 같은 서



[그림 3] 경명 선택을 위한 경명 색인창 화면

비스제공으로 사용자는 보다 편리하게 이경명이나 약경명으로도 검색할 수 있다.

[그림 4] 전체 경을 대상으로 '보살'을 검색한 화면

(2) 용어검색

본 검색 시스템에서는 용어검색을 위하여 각 경전별로 자주 검색하는 용

어, 약 5만여 단어를 선정하여 빠른 검색이 가능하도록 색인화 하였다. 또한 사용자의 다양한 검색 요구에 맞도록 단순히 용어만을 입력받아 전체 검색을 하는 기본검색과 다양한 옵션을 추가하여 검색하는 상세검색 기능으로 나뉘어져 있다. 용어만을 입력하고 검색하면 〔그림 4〕와 같이 기본검색 기능을 수행하게 되는 것으로 전체 경에서 해당되는 용어를 모두 검색한다.

경 선택 후 용어를 입력하면 상세검색을 하게 되는데 이때 '한자검색', '결과내 검색', '히스토리검색', '자연어검색', 그리고 '문장검색' 등의 검색 옵션을 지원하고 있다. 상세검색은 크게 '입력검색'과 '목록검색'으로 나눌 수 있다. 먼저, '입력검색'은 경전을 선택한 후 '용어입력' 상자에 검색할 용어를 직접 입력하여 검색한다. 각 검색 방법에 대한 자세한 설명은 〔표 1〕에 나타내었다.

[표 1] 다양한 용어 검색 방법 및 기능

검색 방법	기능
결과내 검색	이미 검색한 결과와 찾고자하는 결과가 관련이 있는 경우, 검색한 결과 내에서의 재검색은 빠르고 정확한 검색을 가능하게 한다.
한자 검색	독음을 이용한 용어 검색은 음은 같지만 한자가 다른 용어가 있기 때문에, 정확한 검색을 위하여 한자를 이용한 한자 검색 방법이 필요하다. '한자 검색' 옵션 버튼을 선택한 후 용어를 입력하면 한자와 한글독음이 함께 저장된 키워드 테이블(keyword_table)에 접근하여 독음에 해당하는 한자를 검색한다.
히스토리 검색	검색한 결과를 유지하여 이용하는 방법으로 웹 브라우저의 히스토리 기능과 동일하게 동작한다. 즉, '이전' 또는 '다음' 버튼을 이용하여 이미 검색한 결과를 신속하게 다시 제공한다. 히스토리 검색은 단독으로 사용할 수도 있지만 '결과내 검색'과 연동하여 사용할 수도 있다.
자연어 검색	문장 검색과 달리 입력된 문장을 형태소 분석을 통하여 검색 가능한 용어를 추출하고 경 단위로 입력한 단어를 유니코드로 변경하고 유니코드의 패턴매칭을 통해 추출된 용어들을 검색한다.
문장 검색	기존의 용어 검색 방법에서는 용어 선택이나 용어 입력에

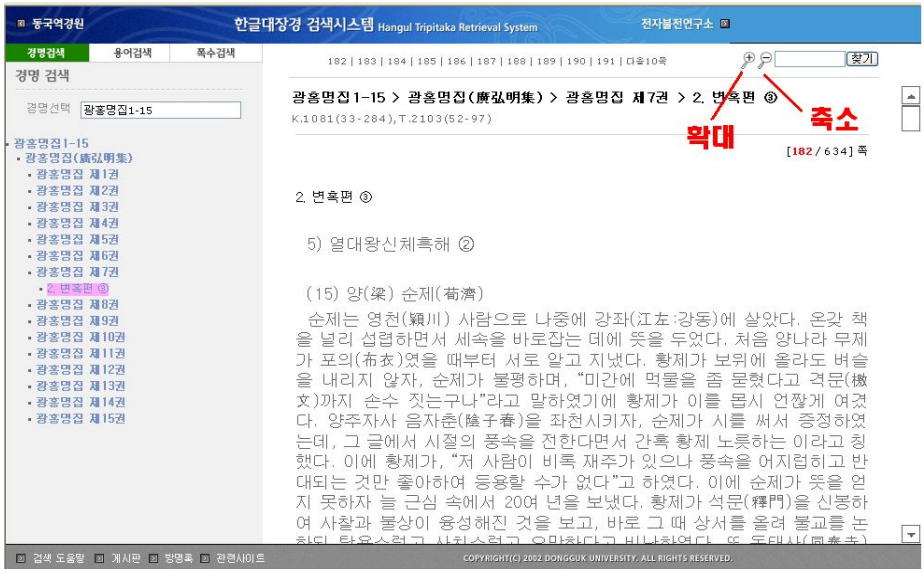
	의한 검색만이 가능 하였지만 문장 검색에서는 입력한 문장 전체가 포함된 페이지를 검색하는 것이 가능하다.

(3) 쪽수검색

쪽수검색은 선택한 경의 쪽을 입력하면, 해당 쪽으로 바로 이동하는 검색 방법이다. 이 방법은 사용자의 페이지 입력 오류를 방지하기 위하여 각 경의 최대 페이지 정보를 자동으로 제공한다. 즉, 사용자가 경을 선택하면 선택된 경의 최대 페이지가 나타나기 때문에 그 페이지를 넘는 검색을 방지할 수 있다. 선택한 경에 따라 본문, 해제, 그리고 서문 정보가 나타나는데, 사용자는 경에 따라 이들 옵션 단추를 선택할 수 있다.

(4) 본문 내용 확대/축소 기능

검색된 본문 화면 크기를 4단계까지 확대할 수 있으며 다시 축소 할 수 있다. [그림 5]와 같이 확대/축소 아이콘을 상단 프레임에 배치하여 사용자가 쉽게 접근할 수 있도록 하였다. 확대/축소 아이콘을 클릭하게 되면 텍스트뿐만 아니라 이미지도 동일한 비율로 확대 혹은 축소된다.



[그림 5] 경명검색에서 선택된 본문 화면

(5) 본문 내 문자열 찾기 기능

검색된 본문 페이지 내에서 원하는 정보를 바로 찾아 볼 수 있도록 문자

열을 입력하면 일치하는 문자열 위치로 이동할 수 있다. 〔그림 6〕과 같이 텍스트 박스에 문자열을 입력 한 후, 찾기 버튼을 클릭 한다. 일치되는 문자열이 있으면 화면이 해당 위치로 이동하고, 일치되는 문자열이 반전된다. 찾기 버튼을 클릭할 때마다 일치하는 문자열로 이동한다.

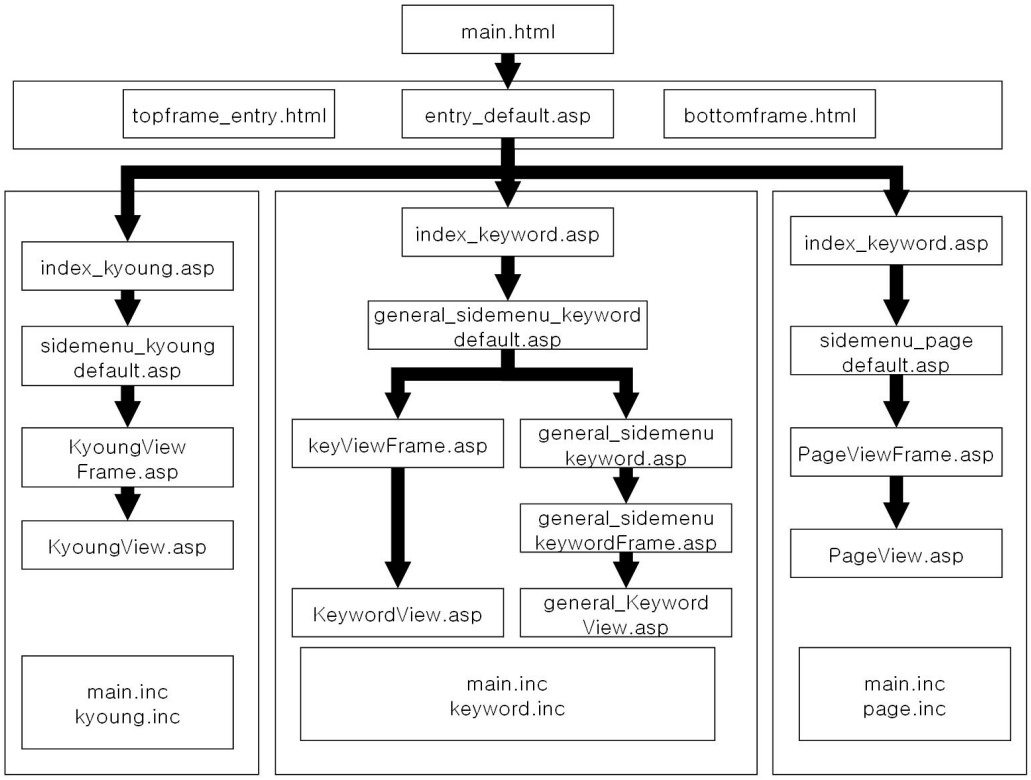
[그림 6] 쪽수검색에서 본문 내 찾기 화면

2.3.2 웹 검색시스템 구현

〔그림 7〕은 한글대장경 웹 검색시스템의 전체적인 제어 흐름을 나타낸 것이다.

한글대장경 웹 검색시스템을 접속하면 main.html이 호출된다. 이것은 세 개의 틀인 topframe-entry.html, entry_default.asp, bottom-frame.html으로 구성되어 있다. 먼저, 상위 틀인 topframe_entry.html은 전자불전연구소와 동국역경원을 링크하도록 구성되어 있다. 다음, 중간 틀인 entry_default.asp는 한글대장경에 대한 간단한 설명과 '검색시작' 단추로 구성되어 있다. 마지막으로, 하위 틀인 bottomframe.html은 검색도움말, 게시판, 방명록, 관련사이트 링크로 구성되어 있다.

사용자가 본 검색시스템의 시작화면에서 검색시작 단추를 누르면, 기본적으로 경명 검색 페이지로 이동한다. 용어 검색과 쪽수 검색으로 이동하기를 원할 때는 해당 탭을 누르면 된다. 사용자가 원하는 검색 결과를 얻는 과정은 다음과 같다. 먼저, 사용자가 해당 검색 페이지에서 검색을 요청하면 본 검색시스템이 사용자의 검색 요청을 질의문으로 변경한다. 그 다음, 원문 저장 데이터베이스에 질의하면 그 결과를 사용자에게 보여준다.



[그림 7] 웹 검색시스템의 제어 흐름도

2.4 누락문자 관리 시스템 확장

누락문자란 현재 윈도우즈 운영체제 및 인터넷 환경에서 사용 가능한 한자에 포함되지 않는 문자를 뜻한다. 이러한 누락문자가 존재하는 이유는 다음과 같이 볼 수 있다.

- 고문헌이 집필될 당시의 한자들이 유니코드 내에 포함되어 있지 않은 경우
- 고문헌의 기록 과정에서 오자 입력으로 인한 실제 존재하지 않는 글자인 경우

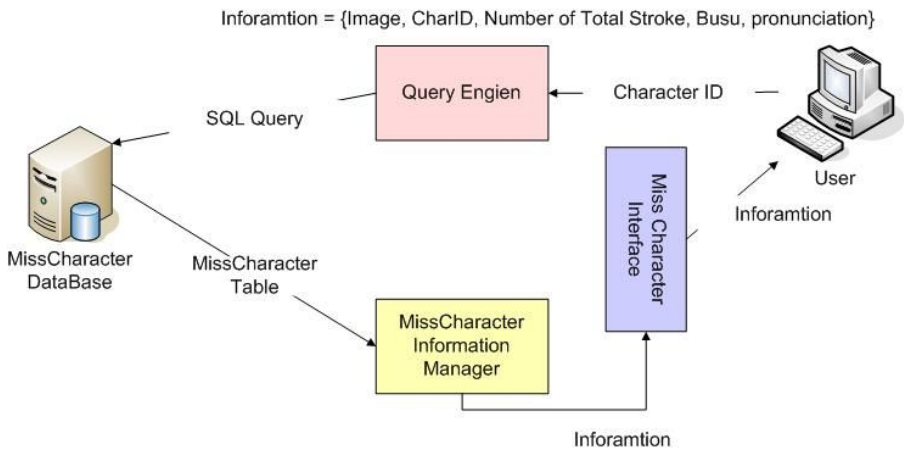
한국 고문헌 상에 나타난 누락문자는 그 자체로 중요한 의미를 갖는다. 이것은 후에 한국 고문헌을 위한 폰트 체계를 정비하는데 있어 도움이 될 뿐만 아니라 한글 대장경 원문 상에 나타난 누락문자의 발생 빈도 등을 한눈에 알아볼 수 있게 해준다. 그리고 누락문자를 사용자에게 서비스하기 위해서는 이미지화 과정을 통해 사용자가 문서를 검색하였을 경우 이질감 및 원문과의 통일성을 부여하여야 한다. 따라서 누락문자 관리기는 문자의 등록 및 이미지 태그 입력 기능뿐만 아니라 각각의 누락문자에 대한 통계자료를 제공하는 기능도 포함하여야 한다.

위의 목표를 달성하기 위하여 몇 년간에 걸쳐 누락문자 생성 및 관리에 필요한 프로그램을 설계하고 구현 하였다. 그리고 앞선 연구에서 기존 누락문자 관리 프로그램을 통해 누락문자 관리의 효율성을 높였다. 그리고 이번 연구에서는 기존 누락문자 관리 프로그램의 기능 확장을 통하여 보다 효율성을 높이고자 한다. 또한, 과거 수년간 축적된 누락문자 이미지들은 생성한 사람들의 전문성 부재로 올바르지 않은 이미지가 존재하는 경우가 있어 정확성을 제공하지 못하고 검색 시스템을 사용하는 사람들에게 가독성을 충분히 제공하지 못하는 경우가 발생할 수 있기 때문에 기존의 이미지를 변경 또는 수정할 수 있는 프로그램을 제작하였다.

2.4.1 누락문자 관리기의 기능 확장

이번 연구에서는 저장되어 있는 데이터베이스에서 하나의 문자를 바로 검색할 수 있는 방법을 제공한다. 이를 “누락 문자 ID를 이용한 검색”이라고 부르고 이는 데이터베이스에 저장된 누락문자 정보를 구분하기 위한 단일키(Primary Key)를 이용한 검색이다. “누락 문자 ID를 이용한 검색”은 만약 누락문자를 처리하는 과정에서 사용자가 누락문자에 대한 문자 ID를 알고 있다면 앞서 세 가지 검색 방법보다 훨씬 빠른 시간 내에 검색이 가능한 장

점이 있다. 일반적으로 한글 대장경의 권별로 누락문자의 출현 빈도를 비교하면 앞 페이지에 나온 누락문자가 연속적으로 나오는 경우가 많음을 알 수 있다. 따라서 매번 같은 문자를 입력하기 위하여 앞선 방법을 이용하는 것은 시간의 낭비를 가져와 업무의 효율성을 떨어뜨릴 수 있다. 앞선 방법으로 누락문자를 검색하여 누락문자의 ID를 기억한다면 새로운 검색 기법은 손쉬운 문자 검색을 제공한다. 아래의 〔그림 8〕은 "누락 문자 ID를 이용한 검색" 구조 및 인터페이스를 보여준다. "누락 문자 ID를 이용한 검색" 구조는 크게 세부분으로 나눌 수 있다. 사용자에게 검색 결과를 보여주는 인터페이스, 누락문자 데이터베이스를 접근하여 해당 누락문자 ID를 가지는 누락 문자에 대한 정보를 가져 오기 위한 SQL Query 문을 작성하는 부분과 데이터베이스에서 전달 받은 누락문자의 정보를 화면에 보여주기 위하여 정보를 관리하는 부분으로 구성된다.



[그림 8] 누락문자 ID를 이용한 검색의 구조도

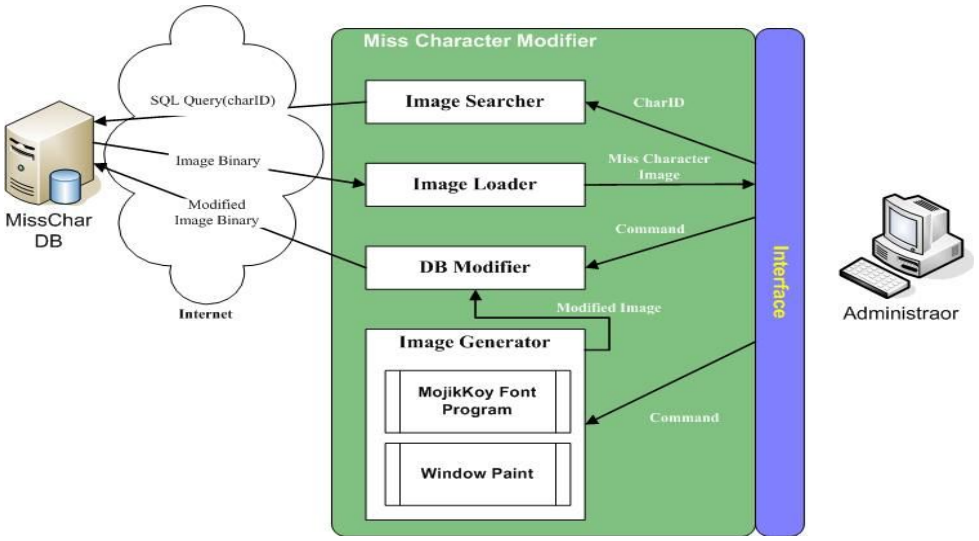
3.4.2 누락문자 수정기

아래의 〔그림 9〕에서 볼 수 있듯이 수년간 누락 문자 처리 작업을 진행해 오면서 발생한 문자의 오류를 수정할 필요가 있다. 문자의 오류 발생 원인은 여러 가지가 있다. 대표적으로 누락문자 처리의 담당자의 전문성 부재로

인한 이미지의 손상 또는 잘못된 이미지 생성이 오류 발생의 첫 번째 원인이라 할 수 있다. 여러 명의 동시 작업에 의한 동기화 문제에 따른 이미지 생성 오류가 두 번째 원인이다. 동기화 문제란 동시에 두 명이 같은 누락 문자를 생성하게 되면 똑같은 문자가 동시에 데이터베이스에 저장됨으로 발생되는 문제를 뜻한다. 마지막으로 문자의 오류가 발생하는 원인은 그림을 생성하는 프로그램인 윈도우에서 제공하는 "그림판"의 한계 때문에 발생한다. "그림판"은 "포토샵"과 같은 전문적인 디자인 프로그램과 달리 프로그램 사용에 많은 전문성을 필요로 하지 않으면서 원하는 간단한 그림을 그릴 수 있는 장점을 지니지만 미세한 그림을 그리는 것에 한계를 가지기 때문에 복잡한 누락 문자의 이미지를 생성하는데 많은 어려움이 따르고 이러한 어려움 때문에 이미지에 오류가 발생한다. 이러한 이미지 오류를 제거하기 위하여 기존에 생성된 이미지를 수정할 수 있는 프로그램이 필요하고 이번 연구에서 누락문자 수정기를 개발하였다.

[그림 9] 누락 문자 이미지 오류 발생

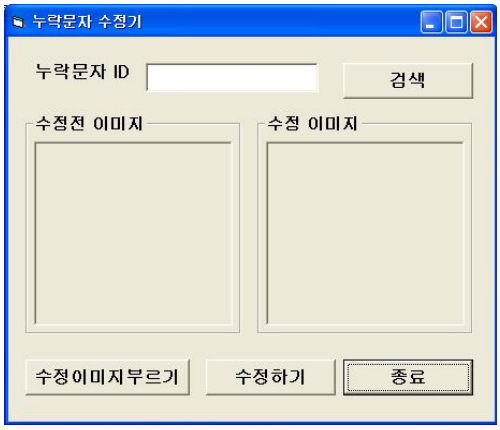
〔그림 10〕은 누락 문자 수정기의 구조를 보여준다. 누락문자 수정기는 크게 사용자 인터페이스와 Database 접근 모듈 그리고 수정 이미지 생성 모듈로 구성된다. 사용자 인터페이스는 오류가 존재하는 누락문자 ID를 입력



[그림 10] 누락문자 수정기 구조도

하는 텍스트 박스, 데이터베이스에 접근하여 이미지를 불러오게 하는 명령 버튼, 기존의 이미지를 화면에 나타내어 사용자에게 보여주는 이미지 박스 그리고 수정된 이미지를 보여주는 이미지 박스로 구성된다. 데이터베이스 접근 모듈은 사용자 인터페이스에서 입력된 누락 문자 ID를 이용하여 데이터베이스에 접근하고 해당 문자에 대한 이미지를 불러오는 역할을 담당한다. 그리고 마지막으로 이미지 생성 모듈은 기존 누락 문자 등록기에서 사용된 모직교 폰트 프로그램과 윈도우의 그림판 프로그램을 실행시키는 역할을 담당한다.

사용자 인터페이스는 〔그림 11〕과 같으며 사용자가 수정할 문자 ID를 입력하고 검색 버튼을 클릭하면 누락문자 데이터베이스에서 해당 문자의 수정 전 이미지가 불리고 수정 이미지를 기다리게 된다. 이미지의 수정은 모직교 폰트 프로그램과 윈도우의 그림판 프로그램을 이용하여 수정을 하고 수정한 후 수정 이미지 부르기 버튼을 클릭하면 수정된 누락문자 이미지가 불려진다. 수정한 이미지를 로드한 후 수정하기를 실행하면 데이터베이스에 수정된 이미지가 저장되고 프로그램은 다른 수정 이미지 ID 입력을 기다리게 된다.



[그림 11] 누락문자 수정기 인터페이스

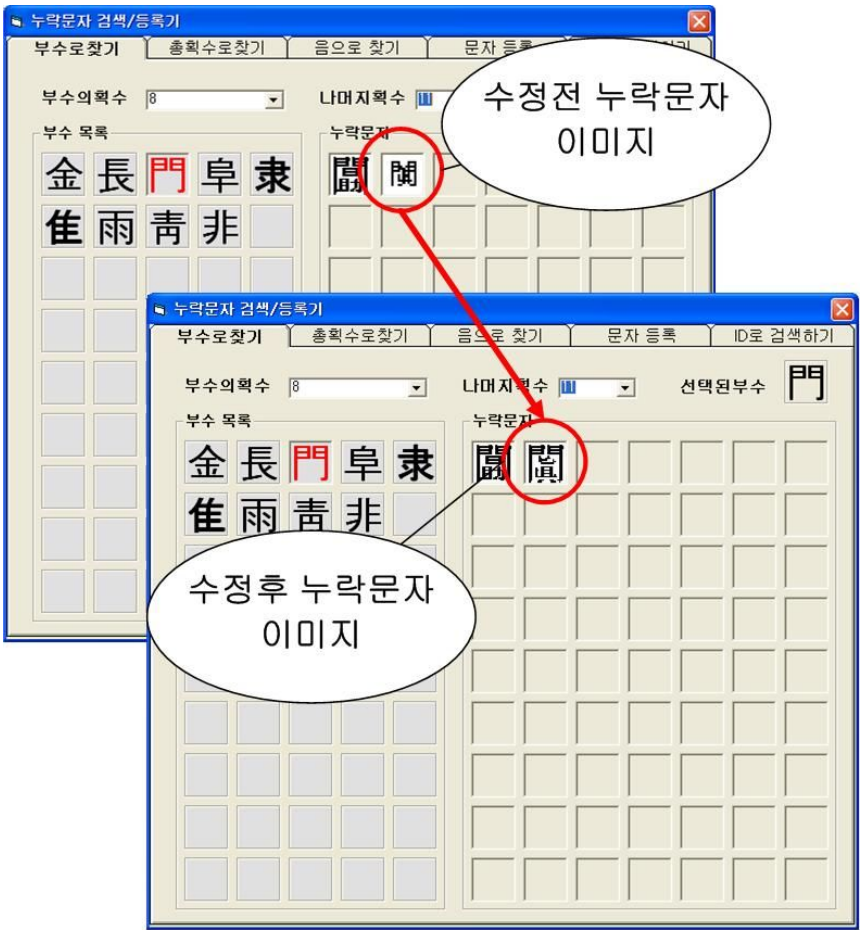
[그림 12] 누락문자 ID "1923" 이미지를 로드한 화면

〔그림 12〕는 수정할 이미지를 불러온 상태를 보여준다. 누락문자 ID가 "1923"인 누락문자를 불러와서 사용자에게 수정을 요청하는 상태를 보여준다. 〔그림 13〕은 오류 이미지를 수정한 화면을 보여준다. 수정 전 이미지와 수정 후 이미지를 비교하면 확연히 사용자의 가독성이 높아 진 것을 알 수 있다.

[그림 13] 수정 이미지를 로드한 화면

위의 과정을 진행한 후 마지막으로 수정하기 버튼을 클릭하면 기존의 잘못된 이미지는 사라지고 새로운 이미지가 데이터베이스에 저장된다. 〔그림

14]는 수정 이미지가 데이터베이스에 저장되어 있음을 보여준다.



[그림 14] 누락문자 데이터베이스의 변경 모습

3. 결론 및 향후 과제

본교는 불교학을 중심으로 한 한국학과 컴퓨터 정보통신 두 분야를 특성화의 큰 축으로 하고 있으며, 불교자료의 전산화야 말로 본교의 특성화 방향인 "불교학과 정보통신 기술"의 연계에 가장 적합한 프로그램이라 할 수 있다. 따라서 수년간 연구를 통해 한국불교전적 중 한글대장경을 전산화하여 본교의 특성화 사업에 부응하고자 하였다.

한글대장경의 전산화를 위하여 가장 필요한 것은 워드프로세서 입력형태로 되어있는 한글대장경 원문을 데이터베이스에 저장하는 기술, 저장된 데이터베이스에서 원하는 부분을 검색하는 기술 및 이를 인터넷에서 사용할 수 있도록 하는 인터페이스 처리 기술이다.

본 연구에서는 한글 워드 프로세서로 작업한 형태의 파일을 일반 유니코드 텍스트로 변환하여 이것을 유니코드 형태 그대로 데이터베이스에 저장하는 기술을 개발 및 구현하였다. 또한 검색 구조를 위하여 문서의 논리적 구조를 표현할 수 있는 XML을 도입하여 재구성 하였으며, 이러한 XML 형태의 문서에서 실제 검색에 필요한 조건들을 추출하여 데이터베이스를 구축하였다.

또한 이렇게 구축된 데이터베이스를 인터넷상에서 열람 및 검색이 가능하도록 웹 기반 프로그램을 작성하였으며, 이를 통하여 인터넷 환경에서 직접 한글대장경을 열람할 수 있도록 하였다. 그리고 여러 가지 검색 기능을 추가하여 사용자가 손쉽게 한글대장경을 열람하고 검색할 수 있도록 하였다. 그리고 유니코드로 표현되지 않는 한자를 인터넷에서 사용할 수 있도록 누락문자를 이미지하고, 원문에 해당 이미지의 URL를 입력할 수 있는 누락문자 관리기를 개발하였다. 누락문자 이미지의 손상 또는 잘못된 이미지 생성에 따른 원문의 이질감을 해소하기 위한 누락문자 수정기를 추가 개발하였다.

본 연구에서 개발된 한글대장경 총 30책 110경이며, 지금까지 총 한글대장경 464경에 대해 인터넷을 통해 검색하고자 한다면 URL "http://ebtc.dongguk.ac.kr"을 이용하면 된다. 향후 연구 과제는 사용자 편이성을 고려 및 사용자의 접근에 대한 통계를 통한 다양하고 편리한 검색 기법의 개발이 필요하다. 그리고 현재 서비스되고 있는 한글대장경의 본문 글자체를 확대 및 축소할 수 있는 기능 추가와 검색 시스템의 화면 인터페이스 변경 등이 필요하다.

참고문헌

[1] Ven. Huimin Bhikkhu, Christian Wittern, and Aming Tu, "CBETA Taisho Electronic Tripitaka," Electronic Buddhist Text, Vol. 3, pp.125-129, 2001.

[2] Ven. Huimin Bhikkhu, Christian Wittern, Aming Tu, Lijuan Guo, and Ray Chou, "A Study on Creation and Application of Electronic Chinese Buddhist Texts: With the Yogācārabhūmi as a Case Study," Electronic Buddhist Text, Vol. 3, pp.49-55, 2001.

[3] Jens Braarvig, "Thesuarus Literaturae Buddhicae (TLB): Its Scope, and a Description of Its Routines," Electronic Buddhist Text, Vol. 3, pp.23-32, 2001.

[4] Dhananjay Chavan, "The Buddha's Words and Electronic Media," Electronic Buddhist Text, Vol. 3, pp.101-123, 2001.

[5] Robert Chilton, "The Asian Classics Input Project (ACIP): Past, Present and Future," Electronic Buddhist Text, Vol. 3, pp.69-88, 2001.

[6] Fred Coulson, "TBRC and Its Model for Linking Text Images with a Bio-Bibliographical Finding Database," Electronic Buddhist Text, Vol. 3, pp.131-145, 2001.

[7] David Germano and Nathaniel Garson, "The Rise of 'Thematic Research Collections' in the Study, Teaching and Transmission of Buddhist Scriptures," Electronic Buddhist Text, Vol. 3, pp.147-190, 2001.

[8] Young Sik Hong, Keum Suk Lee, Yong Kyu Lee, and Tae Sik Han, "Searching Missing Characters from the Hanguk Bulgyo Chonso Database," Electronic Buddhist Text, Vol. 3, pp.253-260, 2001.

[9] C.C. Hsieh, Christian Wittern, and John Lehman, "A Project for Dealing with the Missing Character Problem," Electronic Buddhist Text, Vol. 3, pp.261-269, 2001.

[10] In Sub Hur, "Report on the Digital Tripitaka Koreana 2001," El

ectronic Buddhist Text, Vol. 3, pp.89-100, 2001.
〔11〕 Jae Sung Kim, "A Model of the Unified Tripitaka: Various Versions of the Saddharmapundarika-sutra Processed by XML," Electronic Buddhist Text, Vol. 3, pp.271-278, 2001.
〔12〕 Ishii Kosei, "Lassifying the Genealogies of Variant Editions in the Chinese Buddhist Corpus: N-gram Based System for Variant Document Comparison and Analysis (NGSV)," Electronic Buddhist Text, Vol. 3, pp.33-47, 2001.
〔13〕 Michel Mohr, "Linking Chan/Seon/Zen Figures and Their Texts: Problems and Developments in the Construction of a Relational Database," Electronic Buddhist Text, Vol. 3, pp.219-238, 2001.
〔14〕 Shigeki Moro, "Complex Spatial Digitization Tasks for the SAT Project," Electronic Buddhist Text, Vol. 3, pp.57-68, 2001.
〔15〕 Charles Muller and Michael Beddow, "Moving into XML Functionality: The Combined Digital Dictionaries of Buddhism and East Asian Literary Terms," Electronic Buddhist Text, Vol. 3, pp.191-218, 2001.
〔16〕 Christian Wittern, "Charting of Unknown Territory: Application of Topic Maps to Chan-Buddhist Chronicles," Electronic Buddhist Text, Vol. 3, pp.239-251, 2001.
〔17〕 Unicode enabling, Microsoft Developer's Network, 1997.
〔18〕 Public Unicode Font, ftp://www.ifcss.org/ftp-pub/software/fonts/unicode.
〔19〕 True Type and Unicode, http://truetype.demon.co.uk:80/unicode.htm.
〔20〕 Urs App, "A Look at the Korean Tripitaka Input Project", http://www.iijnet.or.jp/iriz/irizhtml/ebit/samsung.htm.
〔21〕 김무봉, "조선시대 간경도감의 역경사업," 전자불전, 제4집, pp.7-53, 2002.
〔22〕 김성철, "『중론』 Śloka의 제작방식과 번역," 전자불전, 제5집, pp.16-36, 2003.
〔23〕 김은중, "한글대장경 간행의 의의와 과제," 전자불전, 제4집, pp.79-104,

2002.
〔24〕 김재성, “고려대장경 전산화 현황-고려·신수 전산본 일자대조 보고를 중심으로,” 전자불전, 제4집, pp.124-154, 2002.
〔25〕 노진홍, 유응구, 박성은, 이용규, 이금석, 홍영식, “한글대장경 전산화,” 전자불전, 제4집, pp.155-192, 2002.
〔26〕 노진홍, 구현우, 유응구, 박성은, 박영희, 이용규, 이금석, 홍영식, 한보광, “한글대장경 전산화 3차 사업의 현황,” 전자불전, 제5집, pp.108-158, 2003.
〔27〕 묘주스님, “한역경전 번역의 개선방향,” 전자불전, 제5집, pp.80-107, 2003.
〔28〕 이금석, 이용규, 홍영식, 한태식, “한글대장경 검색시스템,” 전자불전, 제4집, pp.105-123, 2002.
〔29〕 전재성, “세계의 현존하는 대장경의 문제점과 일상용어로의 번역,” 전자불전, 제5집, pp.37-62, 2003.
〔30〕 한보광, “일제시대 삼장역회의 성립과 역할,” 전자불전, 제4집, pp.54-78, 2002.
〔31〕 허인섭, “전산화본 고려대장경 2000 완성의 학술적 의미와 미래전망,” 전자불전, 제2집, pp.95-120, 2000.
〔32〕 허일범, “티베트 대장경 번역의 문제점,” 전자불전, 제5집, pp.63-79, 2003.

키워드(Keyword)

한글대장경, 한글대장경 검색 시스템, 한글 대장경 전산화, 유니코드, XML
Hangul Tripitaka, Hangul Tripitaka Retrieval System, Hangul Tripitaka Digitalization, Unicode, XML