한글대장경 전산화 4차 사업

박성은*, 노진홍*, 김영희*, 이재수**,
이용규*, 이금석*, 홍영식*, 한보광***

*동국대학교 컴퓨터공학과
**동국대학교 불교학과
***동국대학교 선학과

요 약

본 연구는 한글대장경 전산화 4차 사업으로 한글대장경 30책 분량을 전산화하여 검색 시스템을 구축하는데 목적이 있다. 고려대장경의 우리말 번역본인 한글대장경을 전산화하기 위해 개역된 고문헌을 입력하여 데이터베이스로 구축하고, 인터넷을 통해 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 웹을 통해 보다 쉽게 전 세계에서 접근할 수 있도록 한다.

I. 서론

본 연구는 한글대장경 전산화 4차 사업으로 한글대장경 30책 분량을 전산화하여 전 세계에서 활발하게 사용되고 있는 인터넷을 통하여 검색할 수 있도록 하는 것이다.

불법이 인도에서 전래되어 인류의 정신문화를 이끌어 온 지도 어언 3000여 년이 지났다. 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, 그 국민들에게 안심낙도의 삶을 제시하였다.

이후 각 나라의 전법승들은 부처님의 가르침을 그 나라의 언어로 전하여 널리 일체중생을 이롭게 하는 역경사업에 진력하였다. 이는 국가의 지원을 받는 경우도 있었고, 전법승만의 불타는 신념에 의한 개인적인 사업인 경우도 있었다. 마침내 전법의 발길이 닿은 국가들에서는 불전을 자국어로 번역하여 편찬·유포하게 되었다.

우리나라에 불교를 전해준 중국에서는 한문(漢文)불경이 편찬·유포된 것이다. 우리나라에서는 중국의 불경을 전해 받아 국가와 국민의 정신적 지주로 삼아왔다. 이는 역사에서도 확인되는 바이다. 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 불

경이 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 정신 문화를 전세계에 알리는 새로운 전법활동이라고 할 수 있다.

II. 한글대장경 전산화 4차 사업

2.1 한글대장경의 입력・교정・색인작업

2.1.1 입력 작업

한글대장경의 입력은 (주)동국전산에 외주를 주어 입력하고, 3차에 걸쳐서 엄밀한 교정작업을 수행하였다. 2004년도 한글대장경 전산화 제4차 사업에서 입력교정한 대장경의 목록은 총 30책, 74경, 496권으로 다음과 같다. (※ K번호는 고려대장경의 경전고유번호임.)

K.22 대보적경(大寶積經)(1-24)
대보적경(25-48)
대보적경(49-72)
대보적경(73-96)
대보적경(97-120)
K.1501 금강삼매경론(金剛三昧經論)
K.1499 종경록(宗鏡錄)(26-49)
K.570 유가사지론(瑜伽師地論)(25-48)

K.949 아비달마품류족론(阿毘達磨品類足論)
K.950 중사분아비담론(衆事分阿毘曇論)
K.1496 부자합집경(父子合集經)
K.30 불설포태경(佛說胞胎經)
K.31 문수사리불토엄정경(文殊師利佛土嚴淨經)
K.1341 대성문수사리보살불찰공덕장엄경(大聖文殊師利菩薩佛刹功德莊嚴經)
K.45 성선주의천자소문경(聖善住意天子所問經)
K.44 불설여환삼매경(佛說如幻三昧經)
K.23 대방광삼계경(大方廣三戒經)
K.28 불설대승십법경(佛說大乘十法經)
K.958 아비담심론경(阿毘曇心論經)
K.959 아비담심론(阿毘曇心論)
K.960 잡아비담심론(雜阿毘曇心論)
K.661 기세인본경(起世因本經)
K.709 불설응법경(佛說應法經)
K.999 불설치의경(佛說治意經)
K.717 불설보법의경(佛說普法義經)
K.708 불설앙굴마경(佛說鴦掘摩經)
K.799 생경(生經)
K.402 대방편불보은경(大方便佛報恩經)
K.125 대승비분다리경(大乘悲分陀利經)
K.1385 대승본생심지관경(大乘本生心地觀經)
K.506 장수왕경(長壽王經)
K.370 금색왕경(金色王經)
K.496 불설묘색왕인연경(佛說妙色王因緣經)
K.488 불설사자소타사왕단육경(佛說師子素馱娑王斷肉經)
K.1469 불설정생왕인연경(佛說頂王因緣經)
K.1173 불설월광보살경(佛說月光菩薩經)

K.212 불설태자모백경(佛說太子慕魄經)
K.210 불설태자묘백경(佛說太子墓魄經)
K.466 불설월명보살경(佛說月明菩薩經)
K.479 불설덕광태자경(佛說德光太子經)
K.1467 불설복력태자인연경(佛說福力太子因緣經)
K.208 불설보살섬자경(佛說菩薩睒子經)
K.209 불설섬자경(佛說睒子經)
K.502 불설사자월불보살경(佛說師子月佛本生經)
K.499 불설대의경(佛說大意經)
K.251 전세삼전경(前世三轉經)
K.252 은색녀경(銀色女經)
K.517 불설과거현재불분위경(佛說過去世佛分衛經)
K.211 불설구색록경(佛說九色鹿經)
K.462 불설녹모경(佛說鹿母經)
K.509 일체지광명선인자심인연불식육경(一切智光明仙人慈心因緣不食肉經)
K.765 수행본기경(修行本起經)
K.775 불태자서응본기경(佛說太子瑞應本起經)
K.112 불설보요경(佛說普曜經)
K.588 섭대승론(攝大乘論)
K.590 섭대승론석(攝大乘論釋)
K.798 선비요법경(禪祕要法經)
K.529 불장경(佛藏經)
K.1263 신화엄경론(新華嚴經論)(1-20)
K.1502 법계도기총수록(法界圖記叢髓錄)
K.1075 속고승전(續高僧傳)(1-10)
K.308 다라니집경(陀羅尼集經)
K.1363 금강정유가문수사리보살공양의궤(金剛頂瑜伽護摩儀軌)
K.1370 유가금강정경석자모품(瑜伽金剛頂經釋字母品)

K.1371 수습반야바라밀보살관행염송의궤(修習般若波羅蜜菩薩觀行念誦儀軌)
K.1373 인왕반야다라니석(仁王般若陀羅尼釋)
K.1506 대방광불화엄경수현분제통지방궤(大方廣佛華嚴經搜玄分齊通智方軌)
K.909 사분율비구계본(四分律比丘戒本)
K.922 담무덕부사분율산본수기갈마(四分律刪補隨機羯磨)
K.1395 근본설일체유부비나야갈치나의사(根本說一切有部毘奈耶羯恥那衣事)
K.1065 대당서역기(大唐西域記)
K.1015 천존설아육왕비유경(天尊說阿育王譬喩經)
K.1017 아육왕경(阿育王經)
K.1018 아육왕식괴목인연경(阿育王息壞目因緣經)
K.1022 찬집삼장급잡장전(撰集三藏及雜藏傳)
K.1504 대장일람집(大藏一覽集)(1-4)
K.1080 홍명집(弘明集)
K.1498 일체경음의(一切經音義)1
일체경음의2
일체경음의3
일체경음의4
일체경음의5
일체경음의6
일체경음의8

2.1.2. 테그 작업

입력과 교정을 마친 30책의 한글대장경은 전자불전연구소에서 페이지・대제목・소제목・해제・서론・각주・진언이미지에 대하여 각각 태깅 작업을 수행하였다.

1) 페이지 태그작업 : 페이지를 검색하여 해당 원문을 보여준다.
2) 제목 태그작업 : 경전의 대제목과 소제목을 검색할 수 있으며, 경전의 제목을 통하여 해당 원문을 확인할 수 있게 한다.
3) 각주 태그작업 : 한글대장경의 각주에 나타나 있는 원문을 확인할 수 있도록 한다.
4) 진언 이미지 태그작업 : 한글대장경에 나타난 진언은 이미지로 처리한 후 이미지로 확인할 수 있도록 한다.

2.2 데이터베이스 저장

한글대장경의 원문은 제목, 원문 내용, 주석 등으로 구성되어있고, 각각의 해당되는 내용은 <JMOK>, <PAGE>, <COMMENT>와 같은 태그들로 구별하기 때문에 이를 이용하여 데이터베이스를 구축할 수가 있다. 이때 원문에 나타나는 한문이 기존 문자 셋으로 표현하는데 한계가 있어서, 원문을 유니코드로 변환하여 저장한다.

먼저, 원문을 데이터베이스에 저장하기 위해서는 각 부분을 구별해주는 태그들이 유효한지 검증하는 작업이 필요하며, 이를 위해 원문을 XML 파일로 저장하여 태그의 유효성을 검증한다. 유효성 검증 작업을 마친 후에는 원문으로부터 제목, 원문 내용, 키워드를 추출하여 유니코드로 변환한 후 그 값을 각 해당 테이블에 저장한다. 즉, 키워드는 키워드 사전으로부터 추출하여 키워드 테이블에 저장하고, 원문은 유니코드로 변환하여 경별로 저장한다. 이러한 한글대장경의 데이터베이스 구축 단계를 간략히 정리하면 다음과 같으며 자세한 내용은 본문을 통해 설명한다.

- 1단계 : 태그가 삽입된 원문의 유효성 검증 작업
- 2단계 : 파일 처리
- 3단계 : 인덱스 추출
- 4단계 : 키워드 및 원문 저장

2.2.1 태그가 삽입된 원문의 유효성 검증 작업

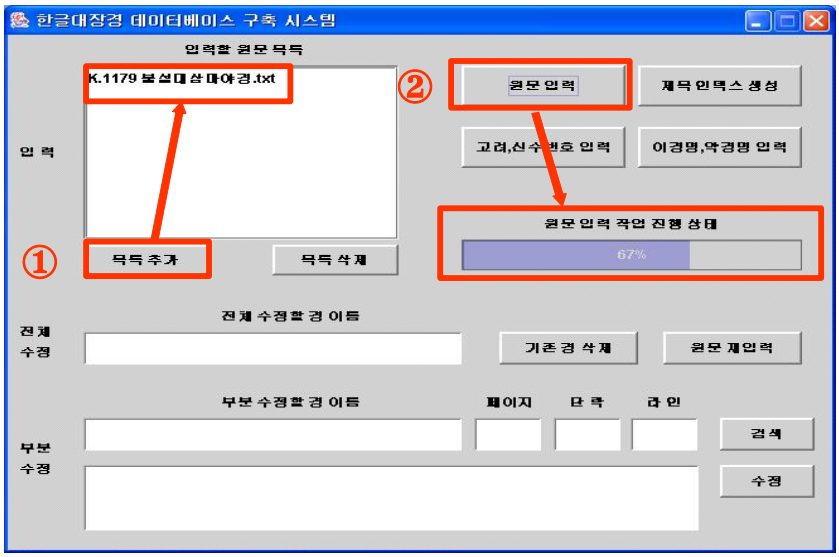
텍스트 파일로 변환된 원문에는 제목, 페이지, 주석, 진언 이미지 등을 구별하기 위하여 각각 <JMOK>, <PAGE>, <COMMENT>, <IMG>라는 태그들을 삽입한다. 이러한 태그들은 여는 태그(<...>)와 닫는 태그(</...>)가 쌍으로 구성되어야하며, 만일 그렇지 않으면 잘못된 데이터가 데이터베이스에 입력될 수 있으므로 반드시 확인 작업을 거쳐야 한다. 이러한 태그들의 검증 작업은 다음과 같은 순서로 이루어진다.

① "*.txt"로 저장된 원문 파일들을 "*.xml"로 확장명을 바꾼다.
② 웹 브라우저에서 해당 XML 문서를 불러들인다.
③ 웹 브라우저에 에러 메시지가 나타나지 않으면 유효한 문서이고, 그렇지 않으면 오류가 발생한 부분을 찾아 원문 내용을 수정해야 한다.
④ 모든 태그들은 여는 태그와 닫는 태그가 쌍으로 이루어져야만 유효한 문서를 생성할 수 있다.
⑤ 최종적으로 이렇게 생성된 유효한 문서를 데이터베이스 구축에 사용한다.

2.2.2 데이터베이스 구축 프로그램의 구현

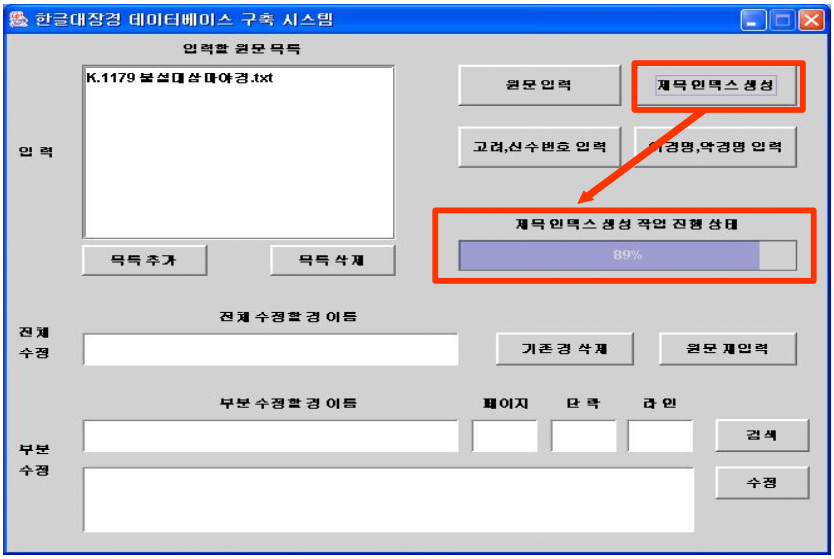
[그림 1]은 데이터베이스 구축 프로그램의 화면이다. 이 프로그램은 한글대장경 원문의 내용을 입력·수정·삭제하는 기능을 지원하며, 원문을 부분적으로 검색하여 수정할 수 있도록 한다.

먼저, 원문을 입력하기 위해서는 왼쪽 중간에 있는 '목록 추가' 버튼을 눌러 '입력할 원문 목록'에 추가한 뒤, '원문 입력' 버튼을 누르면 해당 원문이 저장된다. 원문이 저장되는 과정은 오른쪽 중간에 있는 '원문 입력 작업 진행 상태' 그래프를 통하여 확인한다.



[그림 1] 원문 입력 화면

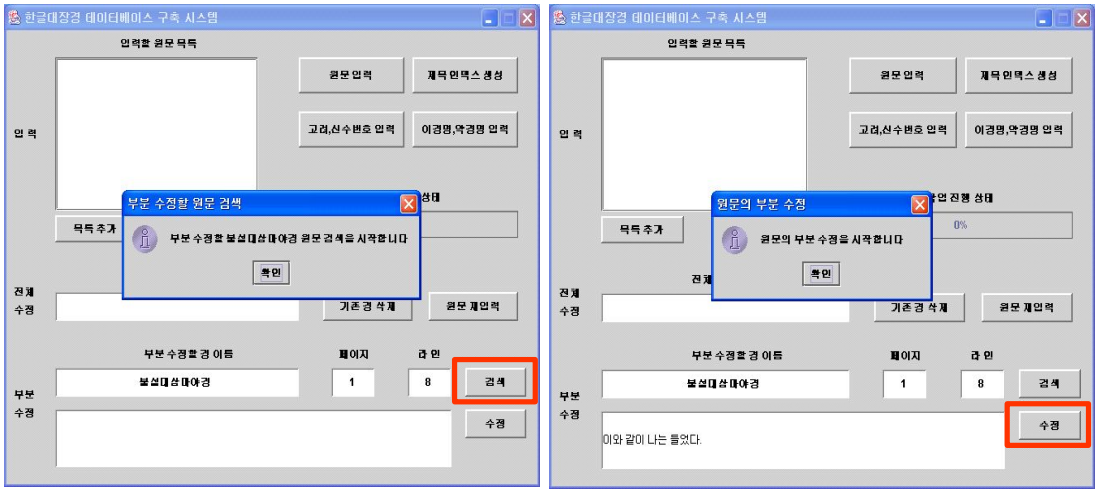
[그림 2]는 데이터베이스 구축 프로그램의 제목 인덱스 생성 화면이다. 오른쪽 상단에 있는 '제목 인덱스 생성' 버튼을 누르면 각 레벨에 해당하는 제목들의 위치 정보를 저장하는 제목 인덱스가 생성된다. 또한 생성된 제목 인덱스 테이블에 '고려,신수번호 입력' 및 '이경명,약경명 입력' 버튼을 누르면 해당 정보들이 입력된다.



[그림 2] 제목 인덱스 생성 및 고려 신수번호 ,이경명 약경명 입력

[그림 3] 전체 수정 화면

[그림 3]은 데이터베이스 구축 프로그램의 원문 전체 수정 화면이다. 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 4] 부분 수정 화면

[그림 4]는 데이터베이스 구축 프로그램의 원문 부분 수정 화면이다. 저장된 내용을 검색하거나 수정할 수 있는 기능이 제공되는데, 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.

2.2.3 원문 파일에서 인덱스 구축

키워드 테이블을 사용해서 원문을 검색하며, 순차검색 방법을 사용하여 검색한다.

(1) 키워드 인덱스 구축

키워드 인덱스를 구축하기 위해서는 키워드 파일에서 순차적으로 키워드를 읽고 각각의 키워드를 원문과 비교하는 방법을 적용했으며, 자세한 과정은 다음과 같다.

① "edocdata"와 "keyword" 테이블을 사용한다.
② 각각의 유니코드로 저장된 테이블의 처음 시작 코드를 읽는다.
③ 키워드 테이블에서 하나의 키워드를 추출하고, 이 키워드를 "edocdata" 테이블의 전체 내용과 비교한다. "keyword" 테이블에서 선택한 키워드와 같은 코드를 발견하면, 일련번호(uid), 키워드번호(keynum), 키워드(keyword), "edocdata" 테이블에서의 페이지(pagenum) 정보를 저장한다. 그리고, 원문 파일의 다음을 비교한다.
④ "eodcdata" 테이블 수만큼 비교한 다음에 다음 키워드를 읽어 앞의 과정을 반복 수행한다. "keyword" 테이블 전체를 비교했을 때, 프로그램을 종료한다.

(2) 제목 인덱스 구축

원문을 검색할 때 <JMOK>과 </JMOK> 키워드를 검색하여 해당되는 부분이 발견되면 각 레벨에 해당하는 제목들의 위치 정보를 “tag_jmok_area_table” 테이블에 저장한다.

<JMOK1> ... </JMOK1>
<JMOK2> ... </JMOK2>
<JMOK3> ... </JMOK3>
<JMOK4 SEARCH='TRUE'> ... </JMOK4>

위에서 보는바와 같이 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.

2.2.4 키워드 및 원문 저장

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.

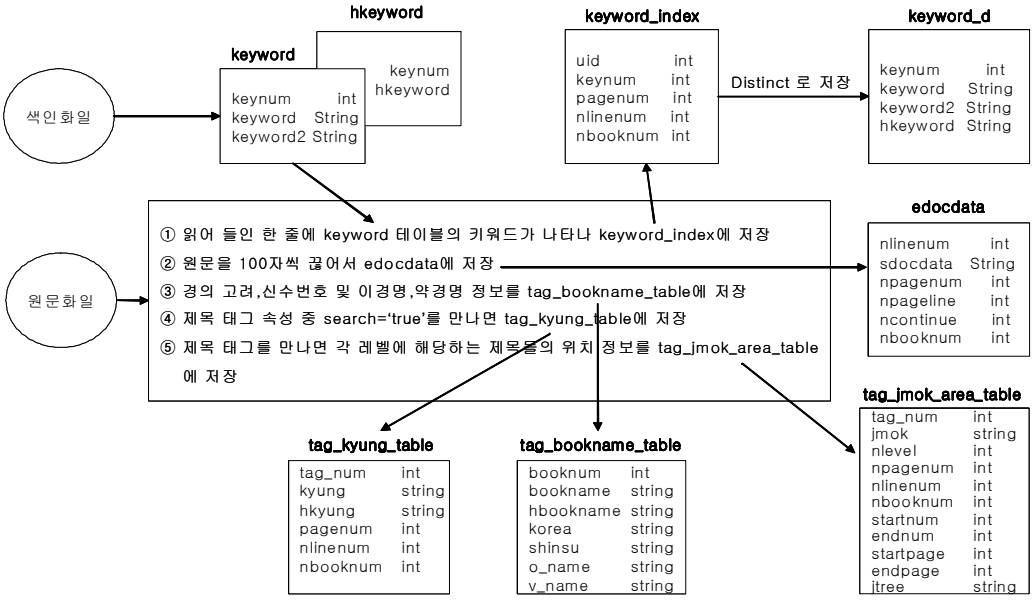
원문 저장은 유니코드 편집기에서 작성된 유니코드 원문을 그대로 테이블에 저장한다. 원문 파일을 라인별로 읽어 저장하면서 페이지 태그를 검사하여 페이지 당 라인(line)수와 “ncontinue” 등의 부가 정보를 생성한다. 원문을 저장할 때 원문에서 한 라인이 레코드의 저장 크기를 초과할 경우에 100자 단위로 나눠서 저장하고, “ncontinue”에 이를 나타내는 부가 정보를 저장한다.

2.2.5 테이블 생성 방법 및 구조 설명

데이터베이스 구축을 위해서 DBMS로는 Microsoft SQL Server 7.0을 사용하며, 본 절에서는 테이블 생성 방법과 주요 테이블 구조에 대한 설명을 한다.

(1) 테이블 생성 방법

[그림 5]는 색인 파일과 원문 파일을 이용하여 각각의 테이블을 생성하는 방법을 나타낸 그림이다. 먼저 색인 파일을 읽어서 “keyword”와 “hkeyword” 테이블을 만든다. 그런 다음 원문에서 읽어 들인 한 줄에 keyword 테이블의 키워드가 존재하면 그 키워드는 “keyword_index” 테이블에 저장한다. 원문은 100자씩 끊어서 “edocdoata” 테이블에 저장하고, 경의 고려·신수번호 및 이경명·약경명 정보는 “tag_bookname_table”에 저장한다. 제목 태그가 나타나면 각 해당 정보를 “tag_kyung_table” 테이블에 저장하고, 제목 태그를 만나면 각 레벨에 해당하는 제목들을 위치 정보를 “tag_jmok_area_table”에 저장한다.

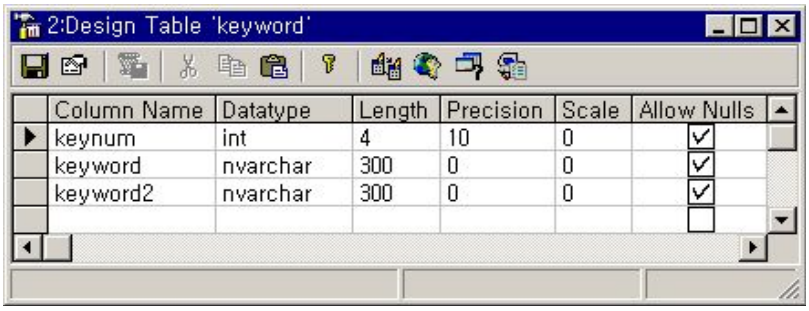


[그림 5] 테이블 생성 방법

(2) 테이블 구조

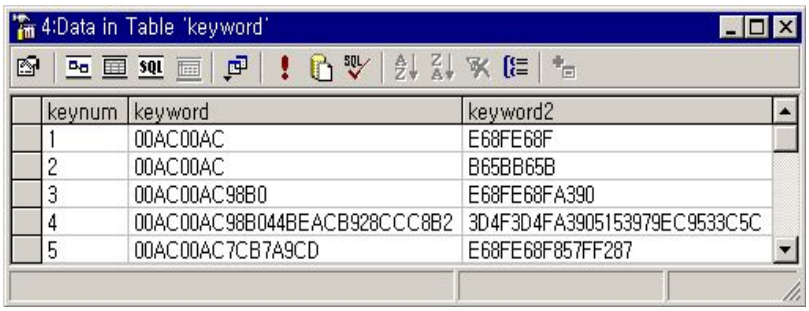
원문의 내용과 인덱스 정보를 저장하고 있는 주요 테이블 구조와 해당 테이블 내용의 예를 살펴본다.

▶ keyword 테이블



Column Name	Datatype	Length	Precision	Scale	Allow Nulls
keynum	int	4	10	0	☑
keyword	nvarchar	300	0	0	☑
keyword2	nvarchar	300	0	0	☑
					☐

[그림 6] keyword 테이블의 레코드 형식

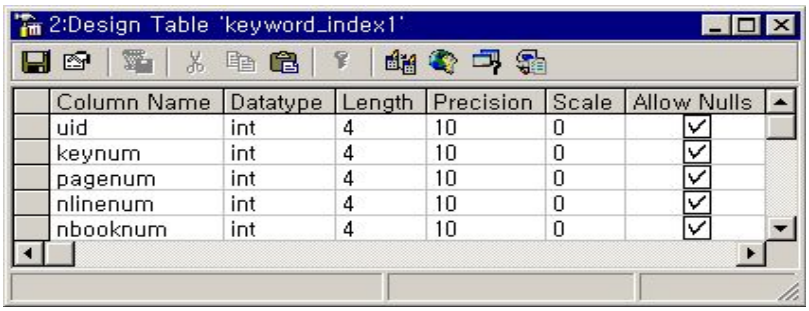


keynum	keyword	keyword2
1	00AC00AC	E68FE68F
2	00AC00AC	B65BB65B
3	00AC00AC98B0	E68FE68FA390
4	00AC00AC98B044BEACB928CCC8B2	3D4F3D4FA3905153979EC9533C5C
5	00AC00AC7CB7A9CD	E68FE68F857FF287

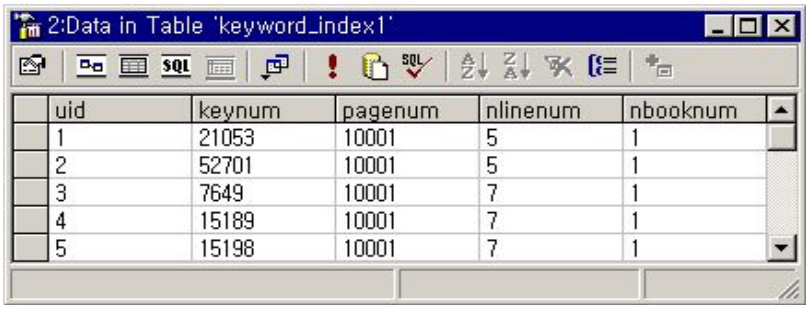
[그림 7] keyword 테이블 내용의 예

① 테이블 명 : keyword
② 테이블 역할 : 키워드 사전으로부터 입력받은 키워드 저장
③ 필드 역할
- keynum : 각 키워드에 대한 유일키 저장
- keyword : 한글 키워드에 대한 유니코드 값 저장
- keyword2 : 한자 키워드에 대한 유니코드 값 저장

▶ keyword_index 테이블



[그림 8] keyword_index 테이블의 레코드 형식



[그림 9] keyword_index 테이블 내용의 예

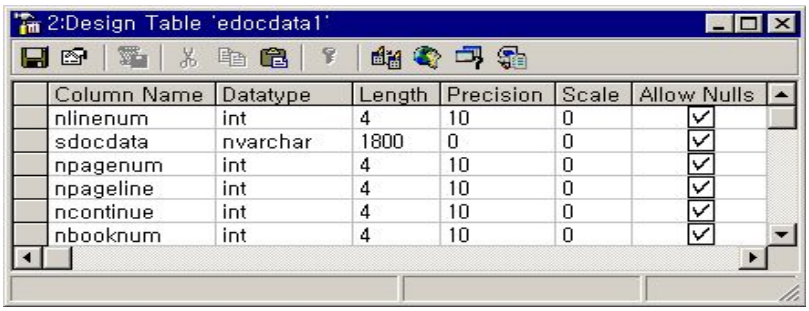
① 테이블 명 : keyword_index

② 테이블 역할 : 각 경별로 키워드 인덱스 테이블을 유지하며, 키워드가 발견된 원문의 경, 페이지, 라인 정보 저장

③ 필드 역할

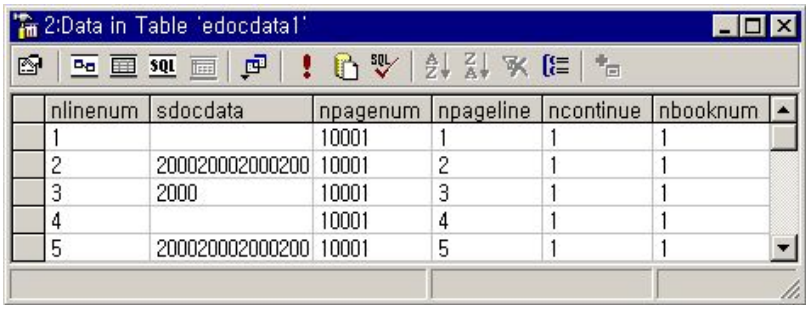
- uid : keyword_index 테이블의 유일키를 저장
- keynum : “keyword” 테이블의 “keynum”과 일치한 값 저장
- pagenum : 키워드가 발견된 곳의 페이지 번호 저장
- linenum : 키워드가 발견된 곳의 라인 번호 저장
- nbooknum : 현재의 경 번호 저장

▶ edocdata 테이블



Column Name	Datatype	Length	Precision	Scale	Allow Nulls
nlinenum	int	4	10	0	☑
sdocdata	nvarchar	1800	0	0	☑
npagenum	int	4	10	0	☑
npageline	int	4	10	0	☑
ncontinue	int	4	10	0	☑
nbooknum	int	4	10	0	☑

[그림 10] edocdata 테이블의 레코드 형식



nlinenum	sdocdata	npagenum	npageline	ncontinue	nbooknum
1		10001	1	1	1
2	200020002000200	10001	2	1	1
3	2000	10001	3	1	1
4		10001	4	1	1
5	200020002000200	10001	5	1	1

[그림 11] edocdata 테이블 내용의 예

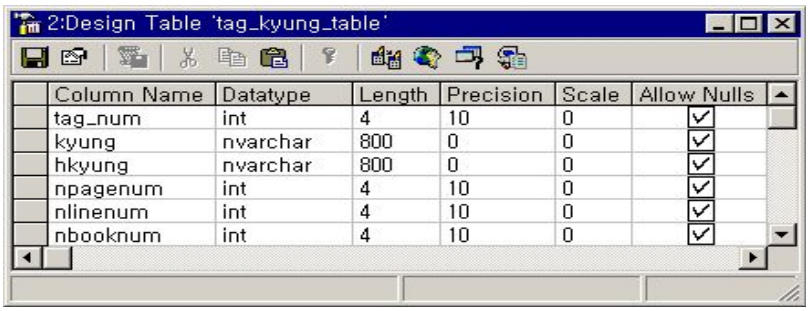
① 테이블 명 : edocdata

② 테이블 역할 : 각 경별로 원문 저장

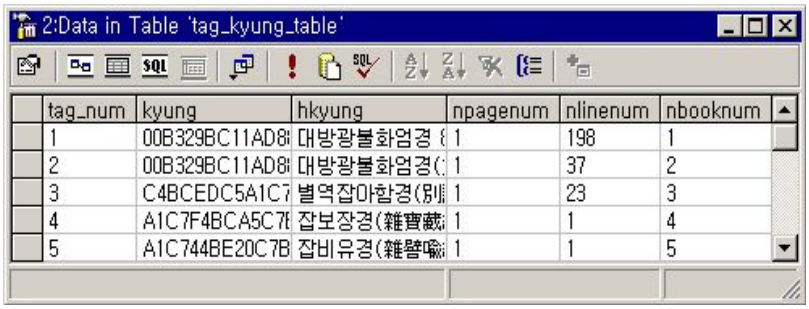
③ 필드 역할

- nlinenum : 원문에 대한 유일키 저장
- sdocdata : 원문을 유니코드 형태로 저장
- npagenum : 페이지 번호 저장
- npageline : 페이지 라인 저장
- ncontinue : 한 라인이 1800자를 초과할 경우 값이 증가
- nbooknum : 현재 경 번호 저장

▶ tag_kyung_table 테이블



[그림 12] tag_kyung_table 테이블의 레코드 형식



[그림 13] tag_kyung_table 테이블 내용의 예

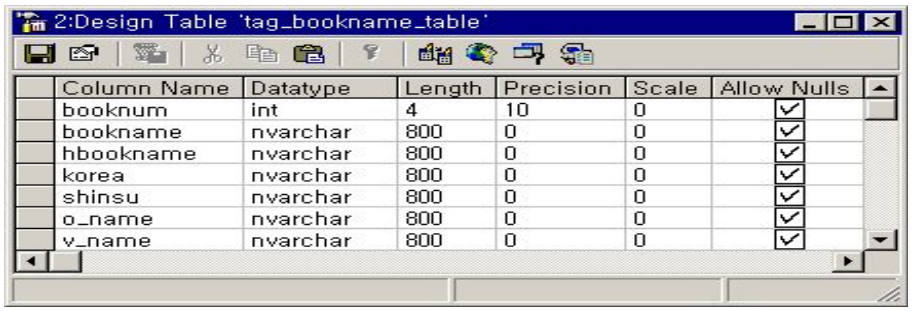
① 테이블 명 : tag_kyung_table

② 테이블 역할 : 경 제목에 대한 정보 저장

③ 필드 역할

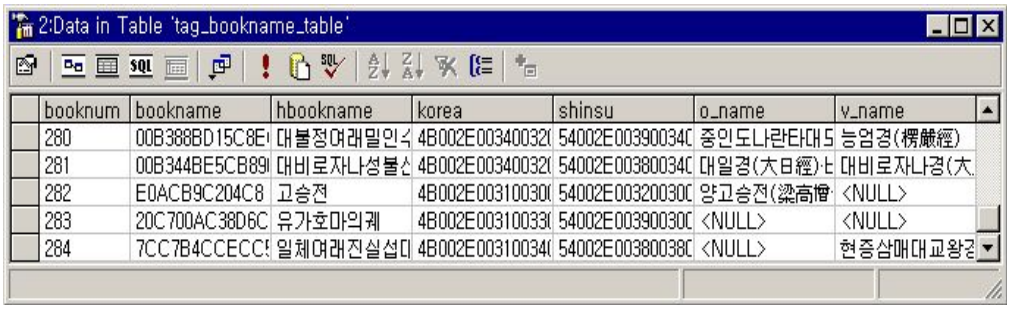
- tag_num : 각 제목에 대한 유일키 저장
- kyung : 경 제목에 대한 유니코드 값 저장
- hkyung : 유니코드에 대한 한글 경 제목 저장
- npagenum : 경 제목이 나타난 곳의 페이지 번호 저장
- nlinenum : 제목의 위치 정보인 “edocdata” 테이블의 “nlinenum”의 값 저장
- nbooknum : 현재 경 번호 저장

▶ tag_bookname_table 테이블



Column Name	Datatype	Length	Precision	Scale	Allow Nulls
booknum	int	4	10	0	☑
bookname	nvarchar	800	0	0	☑
hbookname	nvarchar	800	0	0	☑
korea	nvarchar	800	0	0	☑
shinsu	nvarchar	800	0	0	☑
o_name	nvarchar	800	0	0	☑
v_name	nvarchar	800	0	0	☑

[그림 14] tag_bookname_table 테이블의 레코드 형식



booknum	bookname	hbookname	korea	shinsu	o_name	v_name
280	00B388BD15C8E	대불정여래밀인수	4B002E003400320	54002E003900340	중인도나란타대도	능엄경(楞嚴經)
281	00B344BE5CB89	대비로자나성불신	4B002E003400320	54002E003800340	대일경(大日經)나	대비로자나경(大
282	E0ACB9C204C8	고승전	4B002E003100300	54002E003200300	양고승전(梁高僧	<NULL>
283	20C700AC38D6C	유가호마의궤	4B002E003100330	54002E003900300	<NULL>	<NULL>
284	7CC7B4CCECC	일체여래진실섭대	4B002E003100340	54002E003800380	<NULL>	현증삼매대교왕경

[그림 15] tag_bookname_table 테이블 내용의 예

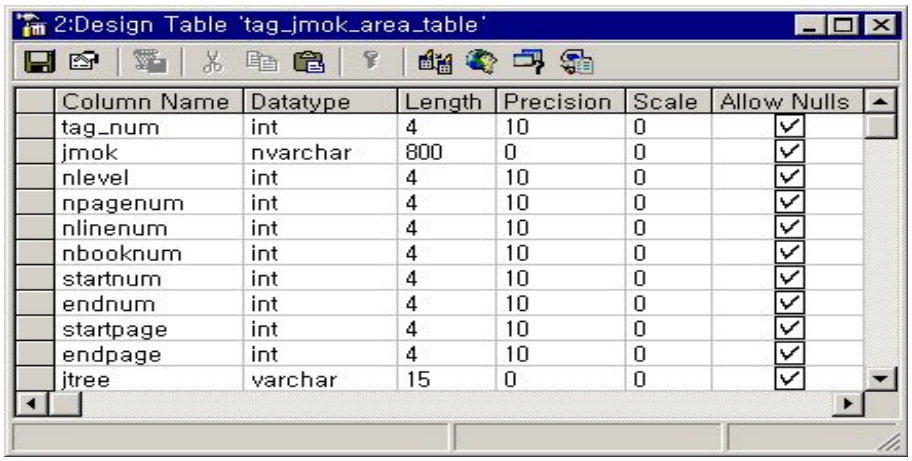
① 테이블 명 : tag_bookname_table

② 테이블 역할 : 각 경이 고려대장경과 신수대장경 어느 부분에 해당되는지 관련 정보를 나타냄

③ 필드 역할

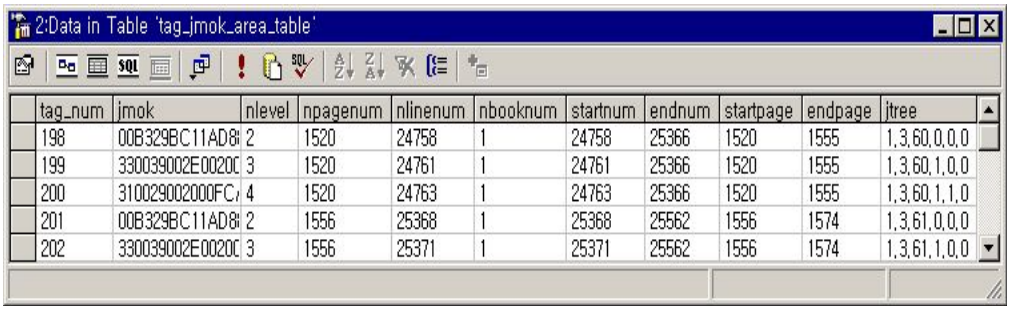
- book_num : 각 경에 대한 유일키 저장
- bookname : 경의 제목에 대한 유니코드 값 저장
- hbookname : 경의 제목의 한글 독음 값 저장
- korea : 고려대장경의 해당 부분 나타냄
- shinsu : 신수대장경의 해당 부분 나타냄
- o_name : 해당 경에 대한 이경명의 정보 저장
- v_name : 해당 경에 대한 약경명의 정보 저장

▶ tag_jmok_area_table



Column Name	Datatype	Length	Precision	Scale	Allow Nulls
tag_num	int	4	10	0	☑
jmok	nvarchar	800	0	0	☑
nlevel	int	4	10	0	☑
npagenum	int	4	10	0	☑
nlinenum	int	4	10	0	☑
nbooknum	int	4	10	0	☑
startnum	int	4	10	0	☑
endnum	int	4	10	0	☑
startpage	int	4	10	0	☑
endpage	int	4	10	0	☑
jtree	varchar	15	0	0	☑

[그림 16] tag_jmok_area_table 테이블의 레코드 형식



tag_num	jmok	nlevel	npagenum	nlinenum	nbooknum	startnum	endnum	startpage	endpage	jtree
198	00B329BC11AD8	2	1520	24758	1	24758	25366	1520	1555	1,3,60,0,0,0
199	330039002E0020	3	1520	24761	1	24761	25366	1520	1555	1,3,60,1,0,0
200	310029002000FC	4	1520	24763	1	24763	25366	1520	1555	1,3,60,1,1,0
201	00B329BC11AD8	2	1556	25368	1	25368	25562	1556	1574	1,3,61,0,0,0
202	330039002E0020	3	1556	25371	1	25371	25562	1556	1574	1,3,61,1,0,0

[그림 17] tag_jmok_area_table 테이블 내용의 예

① 테이블 명 : tag_jmok_area_table

② 테이블 역할 : 각 레벨에 해당하는 제목들의 위치 정보 저장

③ 필드 역할

- tag_num : 각 제목에 대한 유일키 값 저장
- jmok : 경의 제목에 대한 유니코드 값 저장
- nlevel : 각 제목의 레벨 정보 저장
- npagenum : 각 제목이 위치한 페이지 정보 저장
- nlinenum : 페이지 안에서 각 제목이 위치한 라인 정보 저장
- nbooknum : 현재 경의 번호 저장

- startnum : 각 제목에 해당되는 원문 내용이 “edocdata” 테이블에서 시작되는 정보 저장
- endnum : 각 제목에 해당되는 원문 내용이 “edocdata” 테이블에서 끝나는 정보 저장
- startpage : 각 제목이 원문에서 시작되는 페이지의 정보 저장
- endpage : 각 제목이 원문에서 끝나는 페이지의 정보 저장
- jtree : 트리 생성시 계층적 구조 정보 저장

2.3 웹 검색시스템

한글대장경 웹 검색 인터페이스는 사용자가 웹을 통하여 한글대장경을 보다 편리하게 검색할 수 있도록 다양한 검색 방법을 제공하고 있다. 검색 방법은 크게 경명 검색, 용어 검색, 쪽수 검색으로 구성되어있다. 각 검색 방법으로 검색한 결과가 한글대장경의 어느 부분에 속하는지 쉽게 알 수 있도록 위치 정보를 제공하고 있다. 또한 한글대장경을 검색한 후 그 페이지에 해당하는 고려대장경과 신수대장경의 권수와 페이지 정보를 제공하고 있다.

제4차 사업에서는 몇 가지 사용자 요구사항을 추가해서 검색 인터페이스의 기능을 개선하였다. 새로운 요구사항에 맞게 개선한 내용은 용어 검색 기능, 소스코드 보안기능, 각 검색 기능간 이전 정보 유지, 경명 검색의 절차 단순화, 그리고 자유게시판, 방명록, 관련사이트의 화면구조 개선 등이 있다. 또한 시작 화면의 검색기능을 삭제하고 한글 대장경을 소개하는 문구로 수정하였다.

웹 검색 인터페이스의 주요 기능과 제4차 사업에서 수행한 자세한 내용은 본문을 통해서 살펴보도록 한다.

2.3.1 웹 검색시스템의 주요 기능

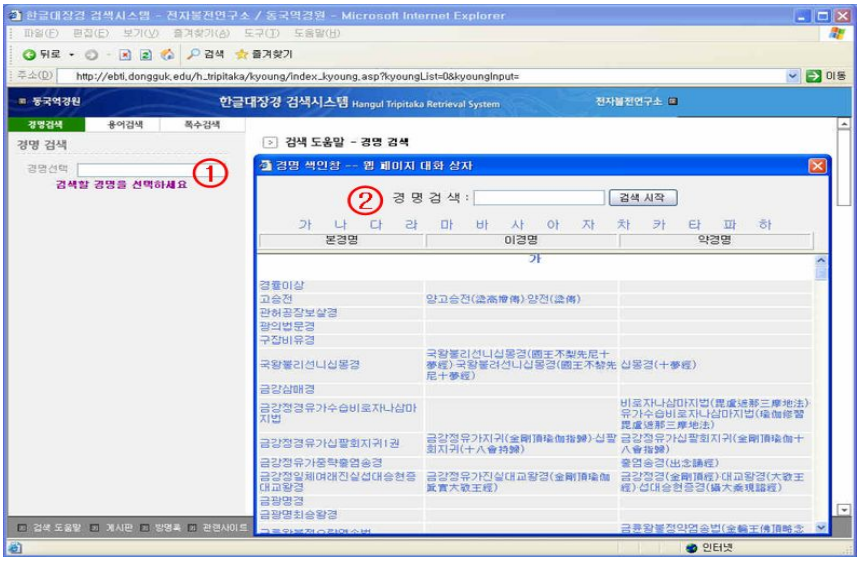
한글대장경 웹 검색인터페이스는 사용자가 편리하게 검색을 할 수 있도록 여러 가지 검색 방법을 제공하고 있다. 주요 검색 방법은 경

명 검색, 용어 검색, 그리고 경명 쪽수 검색 등이 있다.

(1) 경명 검색

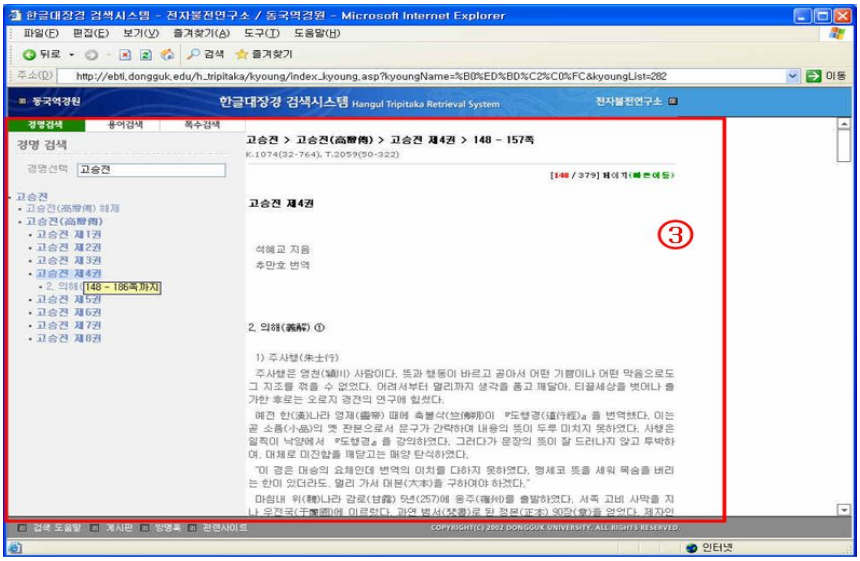
한글대장경의 많은 경전들은 본경명(本經名) 뿐만 아니라 이경명(異經名)과 약경명(略經名)도 가지고 있다. 본경명은 대장경을 잘 아는 전문가 위주의 경명이고, 이경명이나 약경명은 일반 사용자에게 친숙한 경명이다. 본 검색시스템은 본경명은 물론 이경명이나 약경명으로도 쉽게 검색할 수 있는 서비스를 제공하고 있다. 사용자는 이 서비스로 보다 편리하게 이경명이나 약경명으로도 검색할 수 있다.

먼저, 경명으로 검색하기 위해서는 '경명선택'의 입력 상자에 마우스 포인터를 놓고 누르면 본경명, 이경명, 그리고 약경명을 표시하는 색인창이 나타난다. 이것에 해당하는 화면은 [그림 18]과 같다. 경명 색인창에서 원하는 '경명'을 마우스로 누르면 경명 색인창이 사라지고, 좌측 틀에 선택한 경의 목록이 나타난다.



[그림 18] 경명 선택을 위한 경명 색인창 화면

좌측 틀에서 원하는 항목을 선택하면, 오른쪽 틀에는 해당 항목의 본문 내용이 나타나는데, [그림 19]는 2번과 3번 과정을 수행한 결과 화면이다.



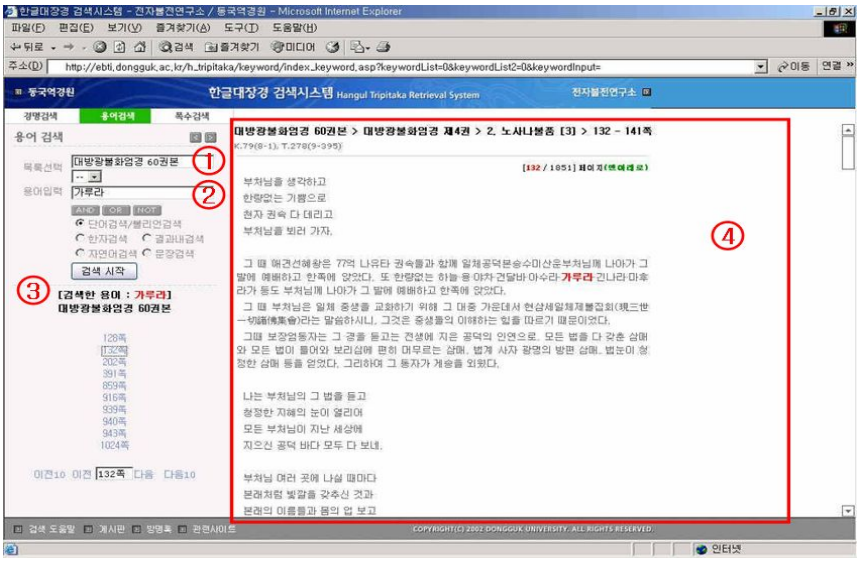
[그림 19] 경명 검색 결과 화면

(2) 용어 검색

본 검색 시스템에서는 용어 검색을 위하여 각 경전별로 자주 검색하는 용어 5만여 단어를 선정하여 빠른 검색이 가능하도록 색인화하였다. 용어 검색은 크게 '입력 검색'과 '목록 검색'으로 나눌 수 있다. 먼저, '입력 검색'은 경전을 선택한 후 '용어 입력' 상자에 검색할 용어를 직접 입력하여 검색한다. [그림 20]은 입력 방법으로 용어를 검색한 화면이고, 그 과정은 다음과 같다.

좌측 틀에 있는 '목록 선택'의 입력 상자에 마우스 포인터를 놓고 누르면 경명 색인창이 나타난다. 이 색인창에서 검색할 경을 선택한다. '용어 입력'의 입력 상자에 검색할 용어를 입력한 후 '검색 시작' 단추를 누른다. 그러면 좌측 틀에는 찾고자하는 용어가 포함된 '쪽'

정보들이 나타난다. 이때 검색한 결과 항목이 많을 경우에는 10개씩 잘라서 보여준다. 좌측 틀에서 원하는 용어가 포함된 '쪽'을 선택하면, 우측 틀에는 선택한 쪽의 본문 내용이 나타난다. 이때 검색한 용어는 빨간색 글씨로 강조되어 나타난다.



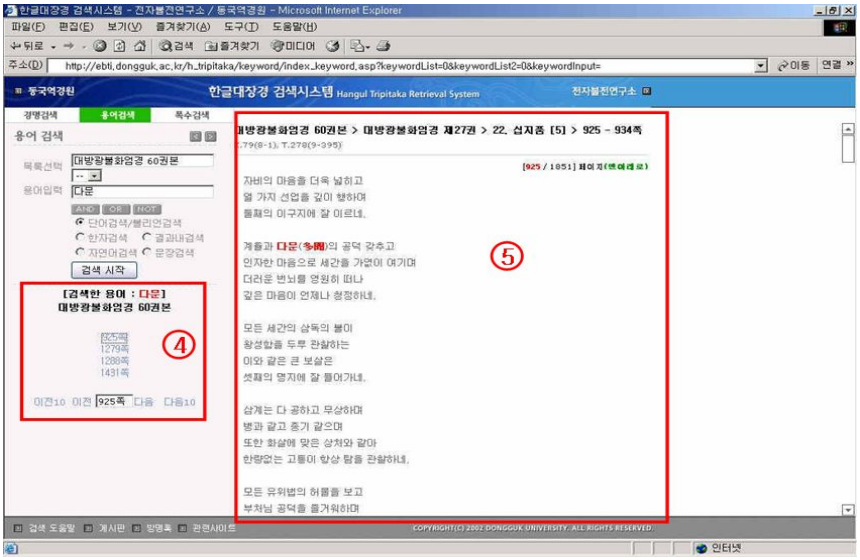
[그림 20] '가루라'라는 용어를 이용한 용어 입력 검색

다음, '목록 검색'은 경전을 선택한 후 한글자음 'ㄱ'부터 'ㅎ'중 하나를 선택하여 검색하는 방식이며 검색 과정은 다음과 같다.

좌측 틀에 있는 '목록 선택'의 입력 상자에 마우스 포인터를 놓고 누르면 경명 색인창이 나타난다. 이 색인창에서 검색할 경명을 선택한다. '목록 선택'의 콤보상자를 누른 후 'ㄱ'부터 'ㅎ'까지 원하는 한글자음 하나를 선택한다. 그런 다음 '검색 시작' 단추를 누르면 선택한 한글자음에 해당하는 용어들이 좌측 틀 하단에 나타난다. [그림 21]은 1번부터 3번까지를 설명한 화면이다.

[그림 21] 목록 선택에서 한글자음 'ㄷ'을 선택한 검색 결과

좌측 틀에서 원하는 용어를 선택하면, 해당 용어가 포함된 '쪽'이 다시 좌측 틀에 나타난다. 이때 검색한 용어들이 많을 경우에는 10개씩 잘라서 보여준다. 원하는 '쪽'을 선택하면, 우측 틀에는 선택한 쪽의 본문 내용이 나타나는데, 검색한 용어는 빨간 글씨로 강조되어 나타나며, [그림 22]는 4번부터 5번까지를 설명한 화면이다.

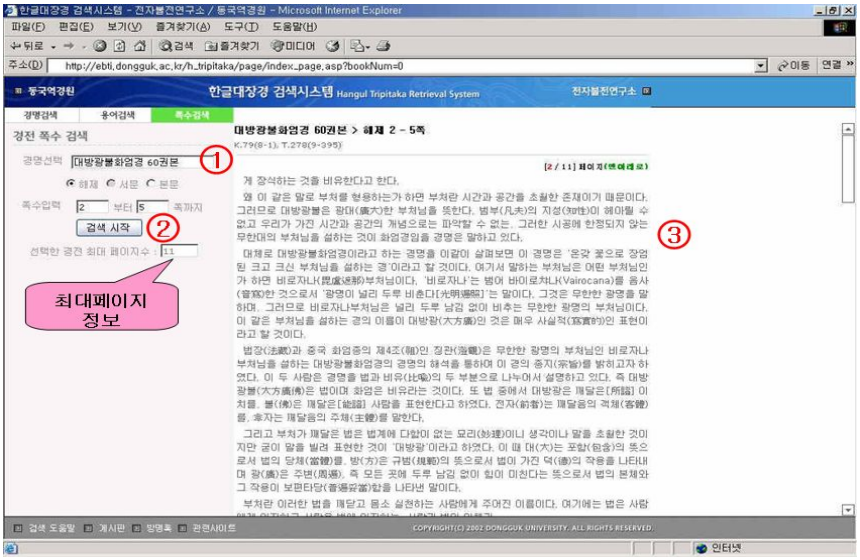


[그림 22] '다문'을 포함하는 쪽을 선택한 화면

(3) 쪽수 검색

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

[그림 23]은 경전의 쪽수 검색하는 화면을 나타낸 것이고, 검색 방법은 다음과 같다. 좌측 틀에 있는 '경명 선택'의 입력 상자에 마우스 포인터를 놓고 누르면 경명 색인창이 나타난다. 이 색인창에서 검색할 경을 선택한다. 이때 선택한 경의 최대 페이지가 좌측 틀에 나타난다. '쪽수' 입력 상자에 찾을 페이지 번호를 입력한 후 '검색시작' 단추를 누른다. 우측 틀에는 선택한 페이지의 본문 내용이 나타난다.



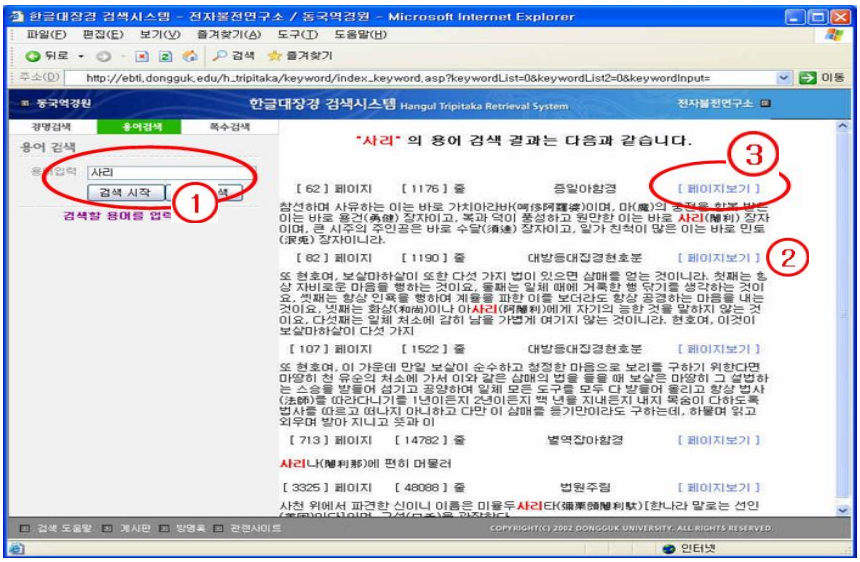
[그림 23] 쪽수 검색의 결과 화면

2.3.2 기능 개선

(1) 용어 검색 기능 개선

기존의 용어 검색은 찾으려는 용어가 어느 경에 있는지를 알아야만 검색이 가능했다. 이 방법은 경의 내용을 잘 알고 있는 불교전문가의 요구사항을 수용한 극히 제한적인 용어 검색이었다. 개선한 용어 검색은 경명을 몰라도 찾으려는 용어를 가지고 전체 경을 검색하도록 하였다. 이것은 일반 사용자의 요구사항을 수용한 검색방법이다. [그림 24]는 개선한 기능으로 용어를 검색한 화면이다. 검색 과정을 보면 다음과 같다.

먼저 좌측 틀의 '용어 입력'의 입력 상자에 검색할 용어를 입력하고 '검색 시작' 단추를 누른다. 우측 틀에는 검색 용어를 포함하는 페이지 정보, 줄 수 정보, 경 이름, 그리고 검색 용어를 포함하고 있는 본문 내용의 일부가 나타난다. '페이지보기'를 누르면 해당 경의 본문 내용이 나타난다.



[그림 24] 용어 검색 기능 개선으로 '사리'용어를 검색한 결과 화면

(2) 소스코드 보안 기능

소스코드란, 프로그램이 어떻게 작동할지를 결정하는 것으로서 소프트웨어 산업에 있어서 중요한 자산이자 기술에 해당한다고 할 수 있다. 본 검색시스템에서도 소스코드가 중요한 기술 자산에 해당한다고 인식을 하였기 때문에 소스코드에 대한 보안기능을 제공하였다.

소스코드에 대한 보안기능을 적용하지 않으면 웹 브라우저를 통해 누구나 손쉽게 소스코드를 볼 수 있다. 그 과정은 웹 브라우저에서 해당 위치에 마우스 포인터를 놓고 오른쪽 버튼을 누르면 [소스 보기] 메뉴가 나타나고, 그 메뉴를 선택하면 해당 부분의 소스코드가 메모장을 통해서 나타난다. 소스코드 보안 기능 적용 후 웹 브라우저에서 손쉽게 소스코드를 볼 수 있는 기능을 차단하였다.

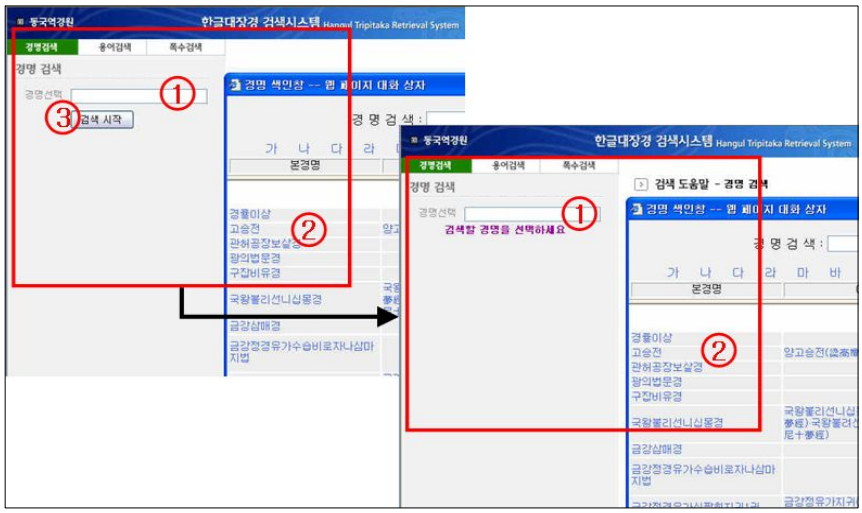
(3) 검색 기능간 이전 정보 유지

기존의 검색 방법은 경명 검색 후 용어나 쪽수 검색으로 넘어가면 유지하고 있던 정보가 사라진다. 그러면 사용자는 처음부터 같은 작업을 반복해야 하는 단점이 있다. 개선한 점은 기존에 검색한 정보가 다른 검색 방법으로 이동해도 이전 정보를 유지하도록 하였다. 이렇게 함으로써 사용자는 새로운 입력을 하지 않아도 되기 때문에 편리한 검색을 할 수 있다. [그림 25]는 '고승전'에 대한 경명 검색 후 쪽수 검색으로 이동했을 때 '고승전'에 대한 정보가 유지되는 것을 보여준다.

[그림 25] '고승전'의 경명 정보의 유지 기능

(4) 경명 검색의 단순화

기존의 경명 검색은 경명 색인창에서 경명을 선택한 후 '검색 시작' 단추를 눌러야만 해당 목록이 좌측 틀에 나타났다. 개선한 점은 경명 색인창에서 경명을 선택하면, 좌측 틀에 선택한 경의 목록을 바로 나타내도록 하였다. 사용자가 '검색 시작' 단추를 누르는 과정을 없앴다. [그림 26]은 기존 경명 검색의 과정과 단순화시킨 경명 검색 과정을 비교한 화면이다.



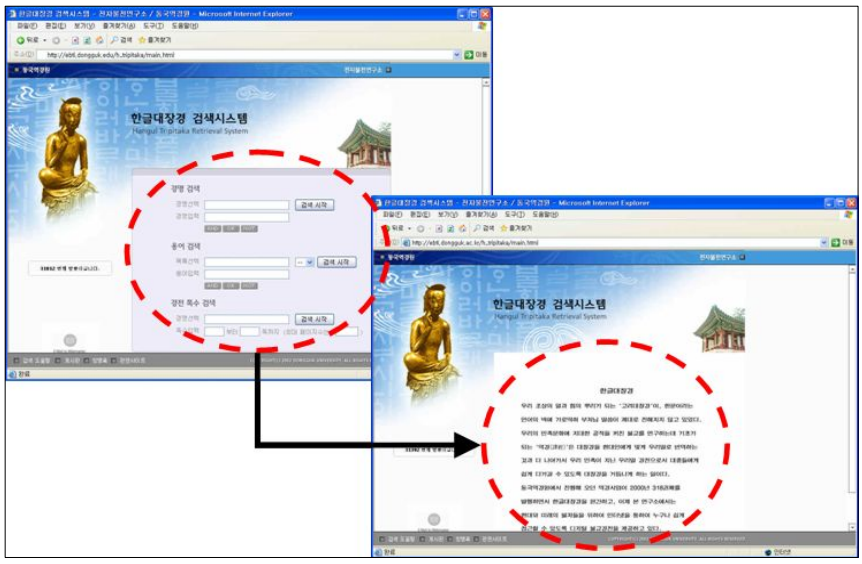
[그림 26] 경명 검색의 단순화

(5) 자유게시판, 방명록 그리고 관련사이트의 화면구조 개선

본 검색시스템의 화면구조는 검색기능을 제공해주는 좌측 틀과 검색 결과 내용을 나타내는 우측 틀로 구성되어 있다. 기존에 제공한 페이지의 구조는 본 검색시스템의 화면구조를 벗어난 형태로 일관성을 해치는 요인이 되었다. 개선한 기능은 본 검색시스템의 기본 구조에 따라 좌측 틀과 우측 틀로 나누어서 일관성을 유지하게 하였다.

(6) 시작화면 개선

시작화면은 사용자에게 목적을 명확하게 전달할 수 있어야 한다. [그림 27]의 좌측은 이전에 제공하였던 화면이고 우측은 개선한 화면이다. 시작 화면의 검색 기능을 삭제하고, 그 위치에 한글대장경을 설명하는 문구로 대치하였다.



[그림 27] 시작화면 중복기능을 제거한 화면

2.4 유니코드에서 누락된 문자 및 진언 처리

누락 문자란, 현재 윈도우즈 운영체제 및 인터넷 환경에서 사용 가능한 한자에 포함되지 않는 문자를 뜻한다. 한글 윈도우즈에서 채택하고 있는 KSC-5601 한글 체계상에서 한자는 대략 4,888자 정도 지원이 되고 있으며, 유니코드를 사용할 경우에는 대략 20,902자 정도의 한자가 지원되고 있다. 그러나 한자로 집필된 불교 고문헌의 경우 KSC-5601 한글 체계나 유니코드 체계에서 지원하지 않는 문자들이 존재하고 있으며 이를 누락 문자(Missing Character)라 칭한다. 이러한 누락 문자가 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

입력 과정을 거쳐야 하는 한글대장경 원문의 분량이 방대하기 때문에 가능한 입력 도구의 간소화와 편리화가 필요하다. 특히, 누락 문자는 수작업을 통해 문서상에 정해진 태그의 형태로 삽입되어야 한다. 따라서 누락 문자 자체를 입력하는 과정이 대단히 번거롭고 시간을 많이 소요하게 되는 작업이 된다. 이러한 누락 문자 입력 과정을 단순화하고 실제 누락 문자를 간단한 방법으로 문서상에 이미지 태그 형태로 삽입할 수 있는 누락 문자 관리기를 개발하였다.

한국 고문헌 상에 나타난 누락 문자는 그 자체로 중요한 의미를 갖는다. 이것은 후에 한국 고문헌을 위한 폰트 체계를 정비하는데 있어 도움이 될 뿐만 아니라 한글대장경 원문 상에 나타난 누락 문자의 발생 빈도 등을 한눈에 알아볼 수 있게 해준다. 따라서 누락 문자 관리기는 문자의 등록 및 이미지 태그 입력 기능뿐만 아니라 각각의 누락 문자에 대한 통계자료를 제공하는 기능도 포함하여야 한다.

2.4.1 누락 문자 관리

유니코드에 없는 문자인 누락문자를 입력하기 위해서는 누락문자를 GIF 형식의 폰트 이미지 파일로 만들고, 누락 문자 DB에 등록한다. 그리고 한글대장경을 인터넷을 통한 검색 시 다른 유니코드 문자들과 함께 등록된 누락 문자를 웹 브라우저 상에서 보여준다.

(1) 원문 입력

한글대장경 원문을 외주를 통해 직접 한글 97 프로그램을 이용하여 직접 입력하고, 이때 입력이 불가능한 한자나 특수 기호는 특별 기호로 표시하게 된다.

(2) 교정 작업

텍스트 파일을 원문과 비교하여 잘못 입력된 내용이 없는지 검토하고 잘못 입력된 내용이 있다면 교정한다. 해당하는 누락 문자의 위치(경, 페이지, 단락, 라인)를 문서화한다.

(3) 누락 문자 입력

교정 작업 도중 누락 문자가 발견되면 누락 문자 검색 프로그램을 사용하여 이미 발견된 누락 문자인지 검색한다. 만약 이미 발견된 누락 문자라면 검색 프로그램을 이용하여 누락문자에 해당하는 Tag를 삽입한다. 여기서 Tag는 누락 문자 이미지가 저장되어 있는 주소를 나타내고 URL 주소로 표현된다. 그러나 저장되어있는 누락문자 중에서 해당 누락 문자를 찾지 못하면 누락 문자를 이미지 파일로 만들고 누락 문자 검색 프로그램에 등록한 후 이에 해당하는 Tag를 삽입한다. 교정 작업 중 다시 누락 문자가 발견된다면 위의 작업을 반복한다.

2.4.2 개선된 누락 문자 관리기(Extended Missing Character Manager)

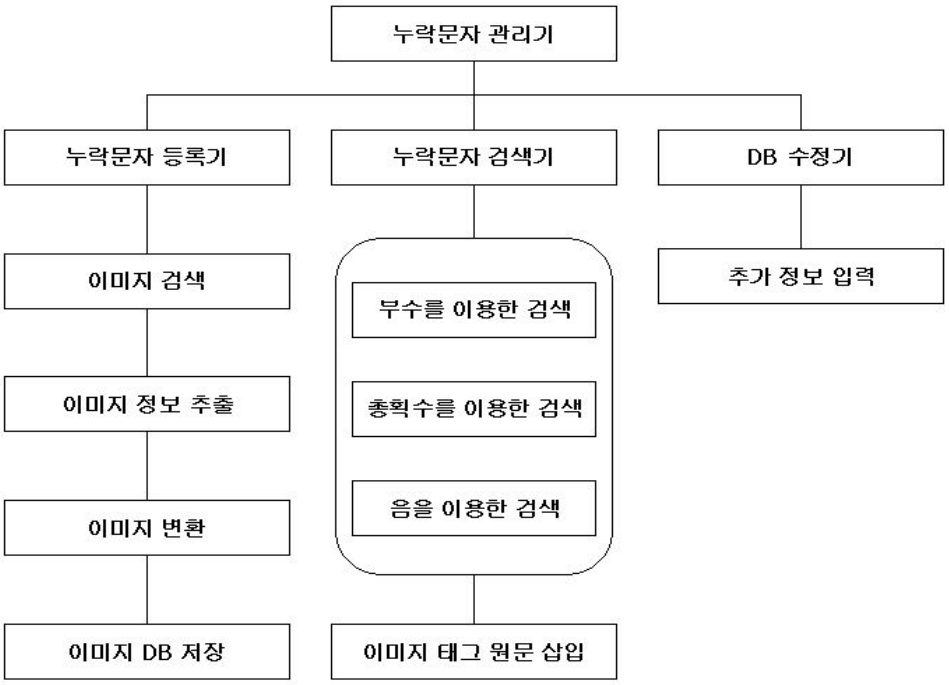
본 과제의 수행에 있어 필요한 누락 문자 관리기의 요구 조건은 다음과 같다.

- 편리한 누락 문자의 등록
- 등록된 누락문자에 대한 빠른 문서상의 입력
- 문서상에 나타난 누락문자의 체계적인 관리 및 통계자료의 제공
- 웹 문서에서 누락 문자 사용의 무 제약성 제공

한글대장경 원문의 입력 작업은 매우 많은 원문의 분량으로 인한 많은 시간이 소요된다. 이런 상황에서 입력과정에서 발견되는 누락문자를 손쉽고 빠르게 등록시킬 수 있는 기능은 필수적이라 할 수 있다. 누락 문자의 등록을 위해서 누락 문자 관리기는 여러 한자 이미지를 체계적인 정렬 방법을 통해 제시하여야 하며, 사용자는 이러한 한자 중에 자신이 찾고자 하는 이미지를 효과적인 방법으로 검색해 낼 수 있어야 한다. 또한 찾고자 하는 누락 문자 이미지가 없는 경우 손쉬운 방법으로 누락 문자 이미지를 작성할 수 있어야 한다.

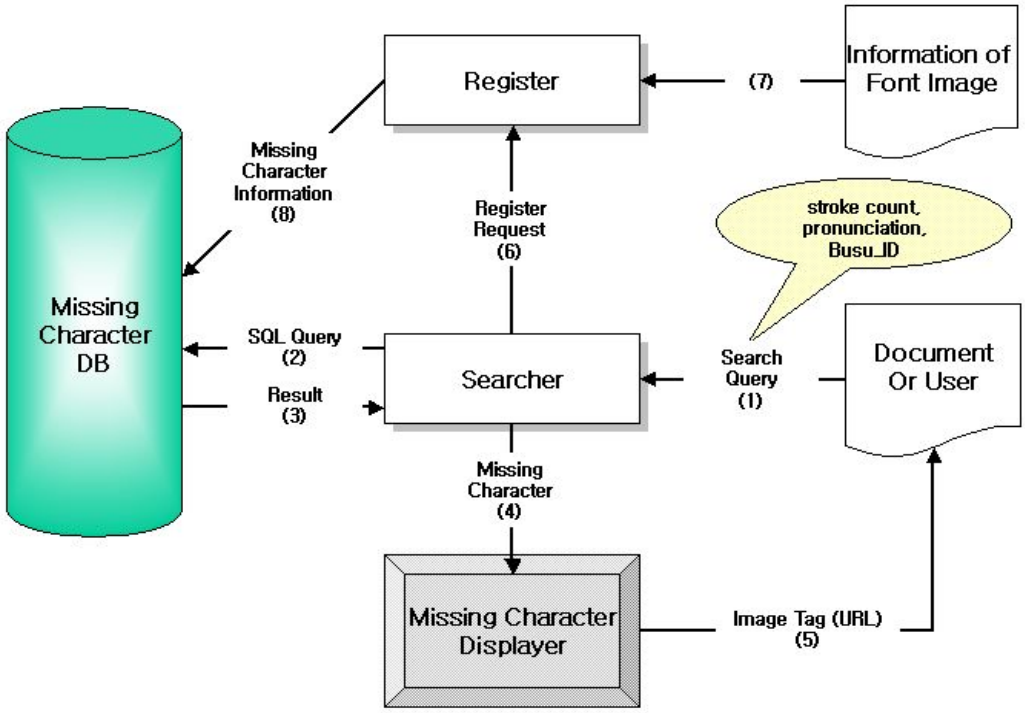
또한 문자 관리기는 등록된 문자에 대해 효과적으로 검색이 가능하여 간편하게 원문 상에 등록된 문자에 대응하는 태그를 입력할 수 있어야 한다. 문자 관리기는 현재 저장하고 있는 등록된 문자의 목록을 효과적으로 게시해 주어야 하며 게시된 문자를 여러 조작 없이, 예를 들어 1회의 마우스 클릭 등의 작업을 통해 원문 상 지정된 위

치에 태그 정보를 삽입할 수 있어야 할 것이다.



[그림 28] 누락 문자 관리기의 구조

[그림 28]과 같이 누락 문자 관리기는 문자의 등록, 검색, 확장전 DB 수정 등 몇 가지 기능별 구조를 갖는다. 문자 등록기 상에서는 표현되지 않는 문자들의 이미지 추출 및 내부적인 코드 부여, DB에 저장하기 위한 이미지의 바이너리 형태로 변환 및 DB에 저장 등 기능을 가지며, 문자 검색기는 현재 등록된 문자의 검색 및 검색된 문자를 원문 상에 삽입하는 기능을 갖는다. 그리고 DB 수정기는 확장전의 DB를 효과적인 검색을 위한 이미지의 추가 정보를 입력하기 위한 인터페이스를 제공한다.



[그림 29] 누락문자 관리기의 동작 과정

[그림 29]는 누락문자 관리기의 전체적인 동작과정을 설명하고 있다. 그림에서 실선은 네트워크로 연결되어 있음을 의미하고 각각의 숫자는 동작 순서를 의미한다.

실질적인 누락문자 관리기에서 가장 먼저 시작되는 과정은 누락문자의 위치정보를 가진 문서에서 누락문자를 검색하기 위한 SQL 쿼리를 검색기에 전달한다. 사용자가 원하는 검색기능을 이용하여 검색기는 네트워크로 연결된 Missing Character DB에서 쿼리에 해당하는 문자의 정보를 가져온다. 이때 가지고 오는 정보(Search Information)는 누락문자 이미지의 태그 정보(URL)와 문자의 ID로 구성된다. 문자의 ID를 데이터베이스에 저장되어 있는 이미지 정보를 임시 이미지 파일로 변환하여 사용자에게 문자 이미지를, 사용자는 보여진 이미지 파일 중 해당 문자가 있는지를 검사하고 만약 해

당 문자가 있으면 원문에 해당 누락문자의 URL을 입력한다. 그렇지 않다면 검색기에서 등록기로 누락문자 등록을 요청한다. 문자 등록기는 누락문자에 대한 이미지와 누락문자 정보를 생성하여 각각 누락문자 디렉토리와 누락문자 데이터베이스에 저장하고 문자 검색기에 저장되었음을 알려준다. 응답을 받은 문자 검색기는 누락문자 검색 과정을 통해 해당 누락문자의 태그 정보(URL)를 원문에 입력한다.

문자 등록기는 한글 대장경 원문 입력 시 나타나는 누락 문자를 문자 관리기에 저장하는 기능을 지니고 있다. 한글대장경 전산화에서는 누락문자를 이미지 형태로 관리하고 있으며 본문 상에 삽입하기 위해서는 HTML의 이미지 태그 정보를 사용하게 된다. 따라서 문자의 등록 과정에서는 필요한 이미지 파일을 작성할 수 있어야 한다.

3차 사업에 사용하였던 누락 문자 관리기는 일반 사용자가 사용하기에 많은 어려움을 가지고 있었기에 4차 사업에서는 누락 문자 관리기의 사용 편이성을 제공하고, 1차에서 3차까지 작성된 누락 문자의 효율적인 관리를 위해 관리기의 개선 작업을 시행하였다. 개선된 누락 문자 관리기에서 검색의 방법을 아래와 같이 여러 기능을 제공함으로써 사용자의 누락 문자 검색을 용이하게 한다.

- 부수를 이용한 검색
- 총획수를 이용한 검색
- 음을 지닌 누락 문자의 음을 이용한 검색

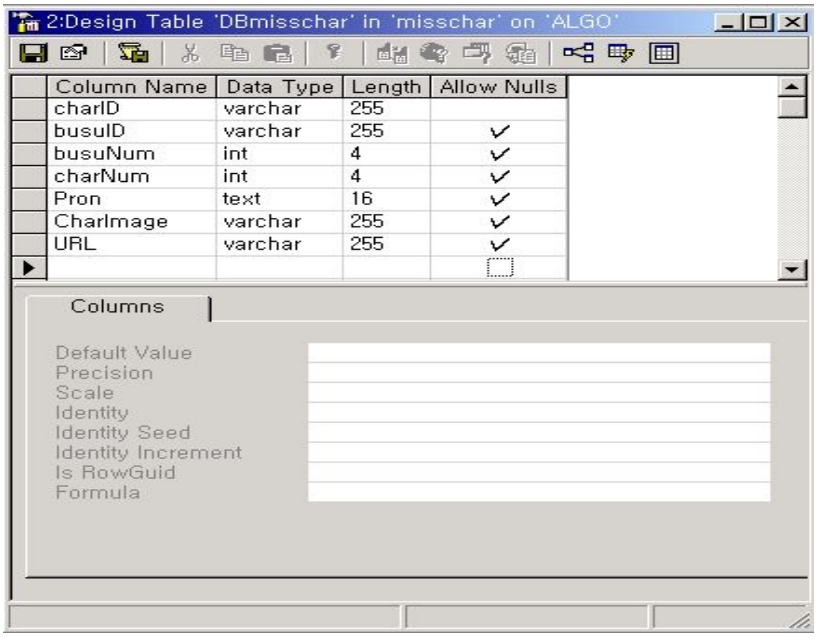
(1) 데이터베이스 확장

여러 가지 검색 기능을 제공하기 위해서는 데이터베이스내의 필드 요소의 추가가 반드시 필요하다. 검색 기능으로 부수를 이용한 검색, 총획수를 이용한 검색, 음을 이용한 검색을 제공하므로 하나의 누락 문자 이미지에 대한 정보를 추가적으로 필요로 하기 때문이다.

	Column Name	Data Type	Length	Allow Nulls
🔑	CharID	char	10	
	CharNum	varchar	10	✓
	url	varchar	100	✓

[그림 30] 기존의 누락 문자 데이터베이스 디자인

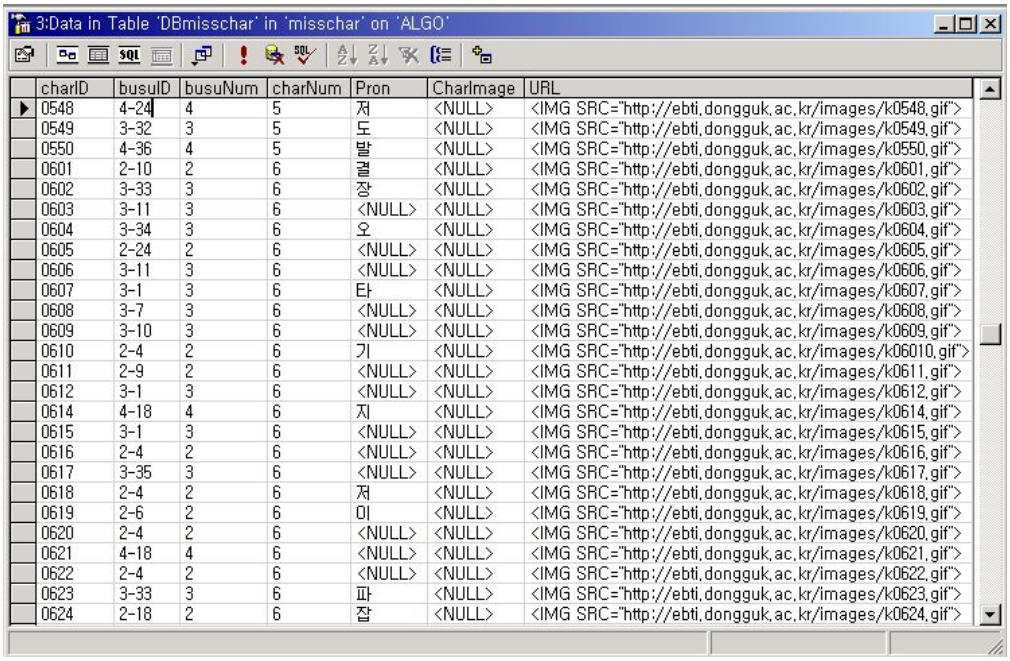
1차에서 3차년도까지 사용했던 누락문자 데이터베이스의 구성 필드는 다음 [그림 30]과 같다. 기존의 누락문자 관리기는 총획수를 이용한 검색 기능만을 제공하였기 때문에 누락 문자의 총획수를 저장하는 필드인 “CharNum”과 해당 누락 문자의 URL을 제공하는 “URL” 필드만 필요하였다.



[그림 31] 개선된 누락 문자 데이터베이스 디자인

[그림 31]은 개선된 누락 문자 데이터베이스 디자인을 보여준다. 기존의 누락 문자 데이터베이스에서 볼 수 없었던 “busuID”,

“busuNum”, “Pron”, “Image”, “Size” 필드가 추가되었다. “busuID” 필드는 한글 2004 또는 MS WORD에서 제공하는 기본적인 부수들을 각각 ID를 주어 해당 부수의 ID를 저장하는 필드이다. 그리고 “busuNum”은 부수의 획수를 뜻하며 “Pron” 필드는 음을 이용한 검색을 위해 누락 문자가 음을 가지고 있다면 해당 음을 저장한다. “Image”는 해당 누락문자의 이미지를 바이너리 형태로 저장하고 “Size”는 이미지의 크기를 바이트 단위로 저장한다.


3:Data in Table 'DBmisschar' in 'misschar' on 'ALGO'

charID	busuID	busuNum	charNum	Pron	CharImage	URL
0548	4-24	4	5	저	<NULL>	<IMG SRC="http://ebti.dongguk.ac.kr/images/k0548.gif">
0549	3-32	3	5	도	<NULL>	<IMG SRC="http://ebti.dongguk.ac.kr/images/k0549.gif">
0550	4-36	4	5	발	<NULL>	<IMG SRC="http://ebti.dongguk.ac.kr/images/k0550.gif">
0601	2-10	2	6	결	<NULL>	<IMG SRC="http://ebti.dongguk.ac.kr/images/k0601.gif">
0602	3-33	3	6	장	<NULL>	<IMG SRC="http://ebti.dongguk.ac.kr/images/k0602.gif">
0603	3-11	3	6	<NULL>	<NULL>	<IMG SRC="http://ebti.dongguk.ac.kr/images/k0603.gif">
0604	3-34	3	6	오	<NULL>	<IMG SRC="http://ebti.dongguk.ac.kr/images/k0604.gif">
0605	2-24	2	6	<NULL>	<NULL>	<IMG SRC="http://ebti.dongguk.ac.kr/images/k0605.gif">
0606	3-11	3	6	<NULL>	<NULL>	<IMG SRC="http://ebti.dongguk.ac.kr/images/k0606.gif">
0607	3-1	3	6	타	<NULL>	<IMG SRC="http://ebti.dongguk.ac.kr/images/k0607.gif">
0608	3-7	3	6	<NULL>	<NULL>	<IMG SRC="http://ebti.dongguk.ac.kr/images/k0608.gif">
0609	3-10	3	6	<NULL>	<NULL>	<IMG SRC="http://ebti.dongguk.ac.kr/images/k0609.gif">
0610	2-4	2	6	기	<NULL>	<IMG SRC="http://ebti.dongguk.ac.kr/images/k0610.gif">
0611	2-9	2	6	<NULL>	<NULL>	<IMG SRC="http://ebti.dongguk.ac.kr/images/k0611.gif">
0612	3-1	3	6	<NULL>	<NULL>	<IMG SRC="http://ebti.dongguk.ac.kr/images/k0612.gif">
0614	4-18	4	6	지	<NULL>	<IMG SRC="http://ebti.dongguk.ac.kr/images/k0614.gif">
0615	3-1	3	6	<NULL>	<NULL>	<IMG SRC="http://ebti.dongguk.ac.kr/images/k0615.gif">
0616	2-4	2	6	<NULL>	<NULL>	<IMG SRC="http://ebti.dongguk.ac.kr/images/k0616.gif">
0617	3-35	3	6	<NULL>	<NULL>	<IMG SRC="http://ebti.dongguk.ac.kr/images/k0617.gif">
0618	2-4	2	6	저	<NULL>	<IMG SRC="http://ebti.dongguk.ac.kr/images/k0618.gif">
0619	2-6	2	6	이	<NULL>	<IMG SRC="http://ebti.dongguk.ac.kr/images/k0619.gif">
0620	2-4	2	6	<NULL>	<NULL>	<IMG SRC="http://ebti.dongguk.ac.kr/images/k0620.gif">
0621	4-18	4	6	<NULL>	<NULL>	<IMG SRC="http://ebti.dongguk.ac.kr/images/k0621.gif">
0622	2-4	2	6	<NULL>	<NULL>	<IMG SRC="http://ebti.dongguk.ac.kr/images/k0622.gif">
0623	3-33	3	6	파	<NULL>	<IMG SRC="http://ebti.dongguk.ac.kr/images/k0623.gif">
0624	2-18	2	6	잡	<NULL>	<IMG SRC="http://ebti.dongguk.ac.kr/images/k0624.gif">

[그림 32] 개선된 데이터베이스의 데이터

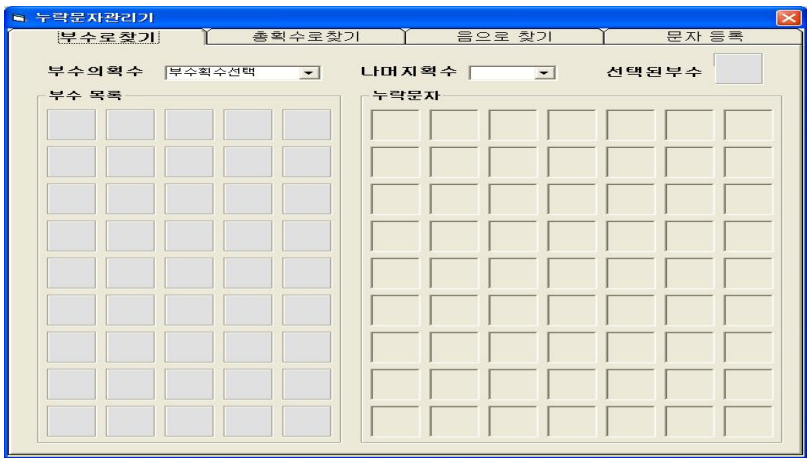
[그림 32]는 개선된 데이터베이스에 데이터들이 입력된 모습을 보여준다. “Pron” 필드의 <NULL> 값은 해당 누락 문자에 음이 존재하지 않음을 보여준다.

(2) 부수를 이용한 검색

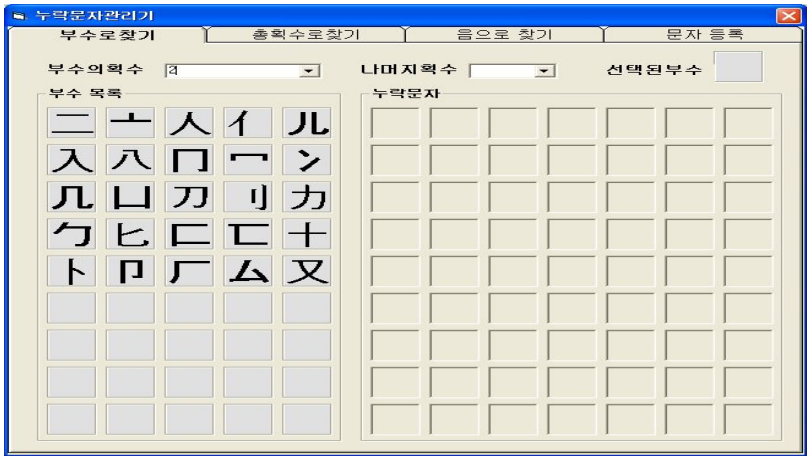
기존의 데이터베이스에 추가된 필드를 이용하여 부수를 이용한 검색을 할 수 있다. 찾고자 하는 누락 문자의 부수를 선택하고 누락 문

자의 총 획수에서 부수의 획수를 제외한 나머지 획수를 선택하면 부수와 나머지 획수를 이용한 SQL 쿼리문이 생성되고 생성된 쿼리 문을 실행시켜 부수와 나머지 획수에 해당하는 누락 문자를 보여준다.

[그림 33]은 부수를 이용한 검색의 초기화면을 나타낸다. 부수의 획수를 콤보 박스에서 선택하면 아래의 [그림 34]와 같이 선택된 부수의 획수를 가지는 부수들이 창의 오른쪽에 보인다. [그림 34]는 부수의 획수를 2로 선택하였을 때 화면이다. 2를 선택하면 부수의 획수를 2로 가지는 부수 25개의 한자가 화면에 나타난다.

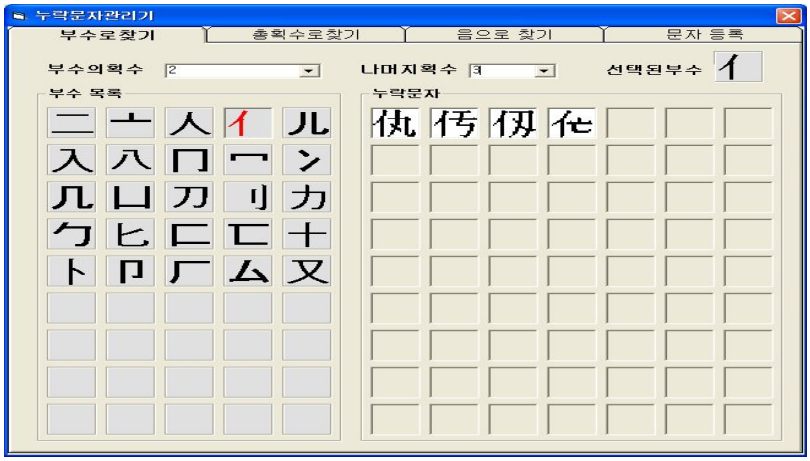


[그림 33] 부수를 이용한 검색의 초기화면



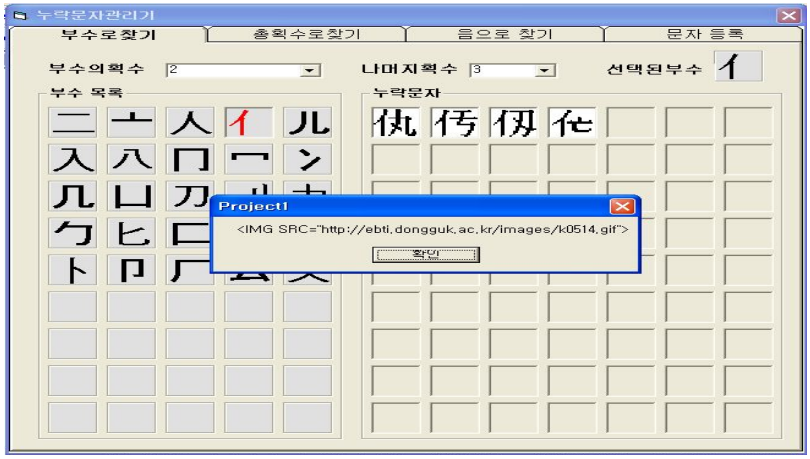
[그림 34] 부수의 획수를 선택한 화면

부수 목록에 나타난 부수들 중에서 찾고자 하는 누락문자의 부수를 선택하면 자동으로 데이터베이스에서 선택된 부수를 가지는 누락문자를 검색하여 나머지 획수를 검색하고 오른쪽 콤보박스에 입력된다. [그림 35]는 부수를 선택하고 나머지 획수를 선택하였을 때 해당 누락문자가 표시되는 화면이다.



[그림 35] 부수와 나머지 획수를 선택 화면

위의 과정을 거쳐 누락문자를 검색한 후 해당 누락 문자를 클릭하면 [그림 36]과 같이 해당 누락 문자의 URL 추출하게 된다.



[그림 36] 누락 문자 선택 화면

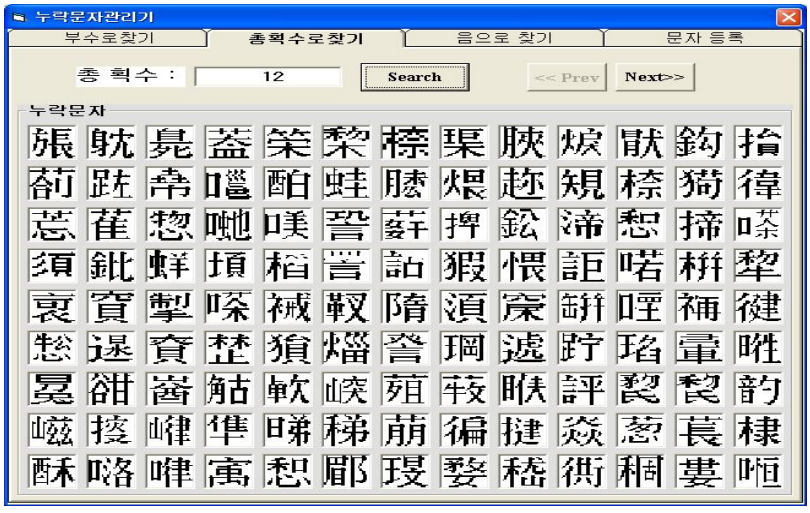
추출된 URL은 클립보드로 누락 문자를 복사한 후 다른 애플리케이션에 복사하기 위해서는 다음과 같은 작업을 한다.

1. 현재 스크린에 실행중인 모든 윈도우들을 검사한다.
2. 윈도우 캡션 이름이 “txt”인 윈도우를 상단으로 불러오고 활성화한다.
3. 복사하기(Ctrl - V) 버튼을 누른다.

위의 3가지 작업을 거치면 원문을 입력 중인 “메모장” 프로그램이 활성화되고, 누락 문자 검색기를 실행하기 직전의 커서가 있던 지점 이후로 누락 문자 이미지 파일에 대한 정보를 가진 이미지 태그가 자동으로 복사된다.

(3) 총획수를 이용한 누락문자 검색

데이터베이스의 “CharNum” 필드를 이용한 총획수 누락문자 검색은 기존의 누락 문자 관리기의 기능과 같다.



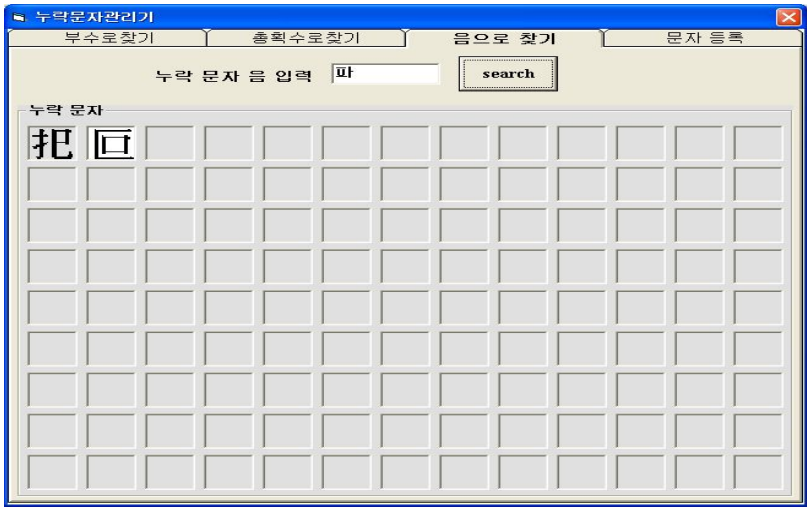
[그림 37] 총획수를 이용한 검색 화면

[그림 37]은 총획수를 이용한 검색 화면을 나타낸다. 한 화면에 나타낼 수 있는 누락문자의 개수가 116개이므로 입력한 총획수를 가지

는 누락문자의 개수가 116개 이상이 되면 Next 버튼이 활성화되어 이후의 누락 문자 이미지 검색이 가능하게 한다. 그리고 Prev 버튼이 이전에 검색했던 이미지로 되돌아가는 기능을 제공한다. URL 추출 과정은 앞서 설명한 부수를 이용한 추출과정과 같다.

(4) 음을 이용한 검색

음을 이용한 검색은 음을 지니고 있는 누락문자만 검색할 수 있는 제한 조건이 있지만 검색되어지는 누락문자의 수가 현저히 적기 때문에 빠른 검색이 가능한 장점을 지니고 있다.



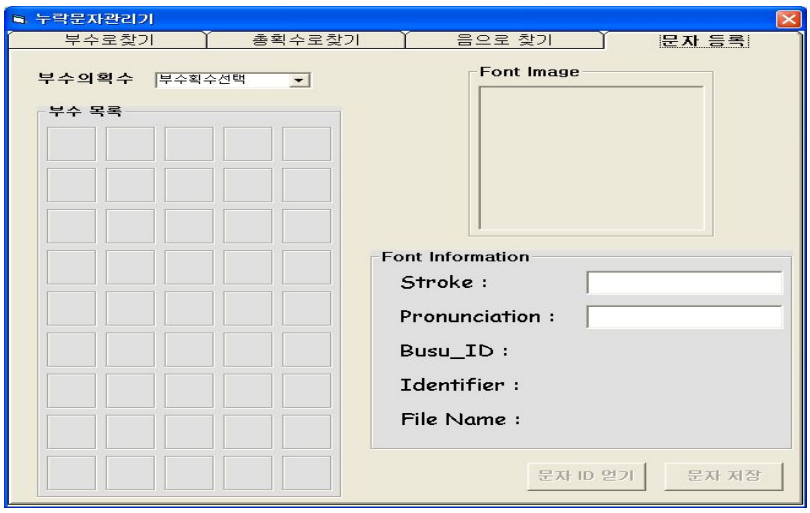
[그림 38] 음을 이용한 검색 화면

[그림 38]은 음을 이용한 검색 화면을 나타낸다. 문자 입력 창에 찾고자 하는 누락 문자의 음을 입력하고 Search 버튼을 누르면 데이터베이스에서 해당 음을 지니고 있는 누락 문자를 검색한 후 이미지를 화면에 출력한다.

(5) 누락 문자 등록

기존의 누락 문자 관리기와 달리 개선된 누락 문자 관리기는 추가된 검색 기능을 제공한다. 추가된 검색 기능을 제공하기 위해 앞서

누락문자의 정보를 저장하는 데이터베이스의 확장이 필요함을 설명하였고, 누락 문자 정보를 추가적으로 입력하기 위해 누락 문자 등록기의 기능도 확장되어야 한다.



[그림 39] 누락 문자 등록기

[그림 39]는 개선된 누락 문자 등록기 화면이다. 기존의 누락 문자 등록기가 모직교 폰트 프로그램을 이용하여 누락 문자 이미지를 생성하였다.

[그림 40] 모직교 폰트의 실행

개선된 누락 문자 등록기도 기존의 누락 문자 등록기와 마찬가지로 모직교 폰트 프로그램을 이용하여 누락 문자 이미지를 생성한다. 모직교 폰트 프로그램을 실행시키기 위해서는 오른쪽 마우스를 클릭하면 나타나는 팝업 메뉴를 이용한다. [그림 40]은 팝업 메뉴를 이용한 모직교 폰트 프로그램의 실행 화면을 타나낸다. 누락 문자를 등록하는 과정은 다음과 같다.

① 모직교 폰트 프로그램을 이용한 누락 문자 이미지 생성
② 부수 선택
③ 총 획수 입력
④ 만약 누락 문자의 음이 존재하면 음을 입력
⑤ 누락 문자 ID 생성
⑥ 데이터베이스에 누락 문자 저장

(6) 누락 문자 이미지 저장

기존 누락 문자 처리기에서는 누락 문자 이미지를 파일로 생성할 때 네트워크 드라이브로 공유된 드라이브에 저장하고, 누락 문자를 검색할 때도 공유된 드라이브에서 이미지를 불러와 작업을 하였다. 네트워크 드라이브로 공유된 드라이브에서 이미지를 저장하거나 불러오는 것이 간단하다는 장점이 있지만, 누락 문자의 수가 많아짐에 따라 네트워크 드라이브로 연결된 드라이브에서 다수의 파일을 불러올 때 시간이 많이 지연되는 단점이 있었다. 그러므로 본 과제에서는 누락 문자에 대한 처리 속도를 향상시키기 위해 누락 문자의 이미지를 데이터베이스에 저장하였다.

누락 문자를 등록할 때에는 데이터베이스의 “Image” 필드에 누락 문자 이미지의 내용을 바이너리 형태로 저장하고, “Size” 필드에는 누락 문자 이미지의 크기를 저장한다. 이렇게 데이터베이스에 저장된 누락 문자 이미지를 불러올 때에는 “Image” 필드의 내용을 바이너리

배열에 저장하고 배열의 내용을 임시 파일에 저장한다. 그리고 임시 파일을 PictureBox 컨트롤로 불러오면 된다.

Ⅲ. 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 본 연구에서는 한국불교전적 중 한글대장경을 전산화하여 본교의 특성화 사업에 부응하고자 하였다.

현재 우리나라에는 귀중한 불교 문헌들을 포함하여 많은 한문 고문헌들이 있으나 이들에 대한 전산화 작업은 아주 미미한 실정이다. 특히 한국불교 및 한문 고문헌에 대한 연구를 하거나, 필요에 의해 한문 고문헌들을 열람하고 싶을 때 귀중한 자료들이 여러 도서관에 분산되어 있어 손쉽게 이용할 수 없다. 그런데 우리나라 불교 문헌 중 고려대장경은 귀중한 문화유산이다. 이러한 고려대장경을 30년에 걸쳐 동국 역경원에서 한글로 번역하여 한글대장경을 완간하였다. 이러한 성과를 이어 한글대장경을 현대 어투로 바꾸고, 번역의 오류를 시정하여 한글대장경 재번역 사업을 시행하고 있다. 따라서 본 연구를 수행하여 전산화하면, 이를 연구하는 연구자들이나 열람을 원하는 사람들에게 도움이 될 뿐만 아니라 우리의 귀중한 문화유산을 전 세계에 널리 알릴 수 있다.

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 데이터베

이스에 저장하는 기술을 개발 및 구현하였다. 또한 검색 구조를 위하여 문서의 논리적 구조를 표현할 수 있는 XML을 도입하여 재구성하였으며, 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.

1차부터 4차사업을 통해 현재까지 개발된 한글대장경 120권본을 인터넷을 통해 검색하고자 한다면 URL "http://ebti.dongguk.ac.kr"을 이용하면 된다. 향후 연구 과제는 확장한자에 대한 처리를 위해 기존에 작업했던 누락문자를 찾아 확장한자를 입력하는 작업이 필요하다. 누락문자가 이미지 파일이기 때문에 화면상 불균형이 생기는 문제를 해결할 수 있다. 그리고 한글대장경의 더욱 많은 부분을 빠른 시일 내에 전산화하는 일이 필요하다. 그리고 데이터베이스에 저장된 내용의 검색을 위해 더 다양한 검색 기법의 도입이 필요하다.

참고 문헌

[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 Pulgyo 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," Electronic 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, 한보광 “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 미래 전망,” 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