

한국불교전서 전산화 5차 사업

구현우*, 박성은*, 노진홍*, 김영희*, 박영희**,
이용규*, 이금석*, 홍영식*, 한보광**

*동국대학교 컴퓨터공학과

**동국대학교 선학과

요 약

현재까지 한국불교전적 및 한국 고문헌에 대한 전산화는 매우 미미한 실정이며, 웹에서의 한국불교전적을 포함한 한국 고문헌에 대한 검색 시스템은 전례를 찾아보기 힘들다. 그러나 문화유산의 전산화 및 전자도서관화는 세계적인 추세이며, 고문헌의 전산화 및 웹에서의 검색 서비스 제공은 해당 분야의 연구자들에게는 시급히 필요한 과제이다. 또한 본교는 불교학을 중심으로 한 한국학과 컴퓨터 정보통신 두 분야를 특성화의 큰 축으로 하고 있으며, 불교자료의 전산화야말로 본교의 특성화 방향인 “불교학과 정보통신 기술”의 연계에 가장 적합한 프로그램이라 할 수 있다.

따라서 본 연구의 목적은 우리의 문화유산인 한국불교전서 중의 9책과 10책을 컴퓨터에 입력하여 데이터베이스에 저장하고, 이를 효율적으로 검색할 수 있는 검색엔진을 개발하여, 현재 전 세계적으로 범용화 되고 있는 인터넷의 웹을 통해 검색할 수 있도록 하는 것이다.

이렇게 본 연구를 수행함으로써 고문헌의 전산화를 활성화할 수 있으며, 전자도서관을 이용한 문화유산의 관리를 촉진할 수 있을 뿐만 아니라, 우리나라 문화유산에 대한 종합적인 전산화 기술을 개발할 수 있다.

I. 서론

본 연구는 한국불교전서 전산화 5차 사업으로 한국불교전서 제 9책과 10책을 전산화하여 전 세계에서 활발하게 사용되고 있는 인터넷을 통하여 검색할 수 있도록 하는 것이다. 한국불교전서는 우리나라에 불교가 전래된 이래 삼국시대부터 우리나라의 선조들이 남긴 옛 문헌들을 발굴 수집하여 출간한 한국불교전서와 해인사 고려대장경을 동국대학교의 역경원에서 한글로 번역한 한글 대장경을 위시한 한국 불교 문헌을 총칭한다.

그 중 한국불교전서는 동국대학교가 30여년 동안의 오랜 시간과 막대한 예산을 투입하여 한국에 불교가 전래된 이래 삼국시대부터 구한말에 이르기까지의 불교계의 고승대덕, 명현학자 등 우리의 선조들이 남긴 옛 문헌들을 낱낱이 발굴 수집하여 전 13책으로 출간한 한국 고전 학술자료의 대총서이다. 이 전서는 신라의 원측이 저술한 ‘반야심경찬’으로부터 구한말의 서진하가 쓴 ‘선문재정록’에 이르기까지 석학 고승 1백 71인이 남긴 2백 88종의 옛 문헌을 그대로 활자화하여 수록한 것인데 대교본을 포함하면 552부 1506권 21편에 이른다. 또한 한국불교전서는 고려의 대각국사 속장경 이후 한국불교의 모든 전적을 집대성한 것으로 우리나라 불교사상의 흐름을 일목요연하게 파악할 수 있을 뿐만 아니라, 정신문화 · 역사 · 철학 등 여러 분야에 걸쳐 효율적인 연구 자료이다. 이의 발간은 우리나라의 역사를 주도해 온 불교사의 심층을 조명하고, 불교 사상과 아울러 한국의 전통사상을 정리한 것이라 할 수 있다[32].

그리고 우리의 귀중한 문화유산인 고려대장경의 한글화 작업이 동국대학교 역경원에서 진행 중이며, 총목록, 목록색인 및 해제, 내용색인 그리고 저역자 색인 등과 번역된 한글 대장경을 2000년 6월까지 총 316여권을 발간하였다[32].

이러한 불교경전에 대한 활발한 편찬 사업에 비하여 불교전적의 전산화 작업은 현재 미비한 상태이나, 세계적으로는 이 분야에 대한 연구가 활발히 진행 중에 있으며 이러한 디지털 경전에 대한 정보교환을 위한 국제 회의도 개최되고 있다[34]. 가상공간에서 불법을 펴고, 법신불 비로자나 부처님을 모시는 방법을 논의하는 국제회의인 전자불전회의(Electronic Buddhist Text Initiative, EBTI)는 ‘전자 경전을 만드는 것을 발의(發意)한다’는 회의다. 불교정보화와 관련 유일한 국제회의인 EBTI는 인터넷에 제공되고 있는 다양한 불교정보를 서로 자유롭게 이용할 수 있는 호환성을 키우는데 그 중요성이 있다. 이 회의를 통해 각국에서 진행하는 불교정보화에 대해 현재까지 진행된 상황 등 여러 가지 의견을 교환한다.

2000년 12월 16일부터 22일까지 미국의 버클리대학에서 20여개국의 70여명의 학자가 참석한 가운데 개최된 EBTI와 2001년 5월 25일부터 26일까지 본교에서 10여개국의 70여명의 학자가 참석한 가운데 개최된 EBTI에서 한문 · 팔리어 · 산스크리트어 · 티벳어 경전에 관한 전산화, 사진, 문헌정보 등이 논의되었고, 불교 문화를 데이터베이스로 저장하는 다양한 프로젝트들도 소개되었다. 뿐만 아니라 전자사전 개발 및 원전 보전 계획도 활발한 것으로 밝혀졌다. 그러나 한문 경전 전산화에서 우리나라에서만 고려대장경이 연구될 뿐, 일본 · 대만 · 중국 · 미국에서는 대정신수대장경에 대한 전산화만을 활발히 추구하고 있는 것으로 나타났다. 또한 EBTI에서는 불경 전산화를 위해 필요한 여러 가지 기술과 표준화도 논의되고 있다. 논의되고 있는 기술 중에서 가장 큰 쟁점이 되고 있는 것은 ‘컴퓨터에서 읽혀지지 않는 한자’ 즉 ‘미싱(missing)’ 한자를 구현하는 방법이다.

이렇게 EBTI에서 논의되고 있는 기술들은 본 연구를 수행하는데도 필요한 기술로서 첫째로 한자를 컴퓨터에 입력할 수 있는 입력방법 및 입력된 한자들을 편집할 수 있는 편집 시스템의 개발이 필요하다.

II. 한국불교전서 전산화 5차 사업

본 연구에서는 한국불교전적을 전산화하여 전 세계에서 이용할 수 있는 인터넷을 통해 손쉽게 검색할 수 있도록 하는데 그 목적이 있다. 이러한 연구 목적을 달성하기 위해 크게 3가지 기술이 필요하다. 한국불교전적을 컴퓨터에 입력하고 이를 편집하여, 데이터베이스에 저장하고, 데이터베이스에 저장된 내용들을 웹에서 검색할 수 있는 인터페이스와 검색 기술이 필요하다. 이들을 위해 본 연구에서 개발한 기술 내용 및 사용방법에 관해 먼저 1절에서는 한국불교전서의 입력 및 교정 방법에 대해 기술하고, 2절에서는 유니코드에 없는 글자를 처리하는 방법, 3절에서는 데이터베이스에 저장하는 방법을 살펴보고, 4절에서는 데이터베이스에 저장되어 있는 내용들을 웹에서 검색할 수 있도록 하는 인터페이스 기술 및 검색방법에 대해 설명한다.

1. 한국불교전서의 입력 및 교정

이번 전자불전연구소에서 수행한 한국불교전서 전산화 5차 사업에서는 한국불교전서 제9책(조선시대편3)과 제10책(조선시대편4)의 전부를 입력하고 교정하였다. 지금까지 한국불교전서 1책·2책·3책·4책·5책·6책·7책·8책·9책·10책을 입력하고 교정, 태그를 달아 전산화함으로써, 한국불교전서의 신라시대편과 고려시대편, 조선시대편의 전산화를 완성하였다. 이는 한국불교전서 전체 분량의 약 80%에 해당하는 분량이다.

제5차 사업에서 입력·교정한 저술은 제9책 총 37종과 제10책 총 49종 모두 86종이다. 입력한 저술들의 목록은 다음과 같다

제 9 책 (조선시대편 3)

雲峯禪師心性論(一卷)	釋大智
月峯集(三卷)	釋策憲
念佛普勸文(一卷)	釋明衍
月渚堂大師集(二卷)	釋道安
楓溪集(三卷)	釋明察
百愚隨筆(一卷)	釋明眼
般若波羅蜜多心經略疏連珠記會編(二卷)	釋明眼
現行法會禮懺儀式(一篇)	釋明眼
禪源諸詮集都序科評(二卷)	釋秋鵬
雪巖雜著(三卷)	釋秋鵬
雪巖禪師亂藁(二卷)	釋秋鵬
無用堂遺稿(二卷)	釋秀演
無竟集(三卷)	釋子秀
無竟室中語錄(二卷)	釋子秀
佛祖眞心禪格抄(一卷)	釋子秀
禪門五宗綱要(一卷)	釋志安
喚惺詩集(一卷)	釋志安
影海大師詩集抄(一卷)	釋若坦
虛靜集(二卷)	釋法宗
禪源集都序科記(二卷)	釋定慧
別集別行錄節要私記解(一卷)	釋定慧
南嶽集(一卷)	釋泰宇
松桂大禪師文集(三卷)	釋瀨湜
霜月大師詩集(一卷)	釋璽筠
天鏡集(三卷)	釋海源
請擇法報恩文(一卷)	釋快善
箕城大師念佛還鄉曲(一卷)	釋快善

月波集(一卷)	釋兌律
龍潭集(一卷)	釋慥冠
新編普勸文(一卷)	釋有璣
好隱集(四卷)	釋有璣
雪潭集(二卷)	釋自優
野雲大禪師文集(三卷)	釋時聖
鰲巖集(一卷)	釋毅旻
龍巖堂遺稿(一卷)	釋體照
大圓集(一卷)	釋大圓

제 10 책 (조선시대편 4)

默庵大師詩抄(三卷)	釋寂訥
諸經會要(一卷)	釋寂訥
秋波集(三卷)	釋泓宥
秋波手柬(一卷)	釋泓宥
西域中華海東佛祖源流(一卷)	釋采永
四溟堂枝派根源錄(一卷)	釋諶諶
三門直指(一卷)	釋捌關
振虛集(二卷)	釋捌關
都序科目并入私記(一卷)	釋有一
法集別行錄節要科目并入私記(一卷)	釋有一
蓮潭大師林下錄(四卷)	釋有一
釋典類解(一卷).....	釋有一
括虛集(二卷)	釋取如
冲虛大師遺集(二卷)	釋旨冊
蒙庵大師文集(二卷)	釋箕穎
大方廣佛華嚴經義湘法師法性偈科註(一卷)	釋有聞

月城集(一卷)	釋費隱
仁嶽集(三卷)	釋義沾
鏡巖集(三卷)	釋應允
三峯集(一卷)	釋知濯
澄月大師詩集(三卷)	釋正訓
大東禪教考(一卷)	丁若鏞
禪文手鏡(一卷)	釋亘璇
修禪結社文科釋(一卷)	釋亘璇
作法龜鑑(二卷)	釋亘璇
無住警策(一卷)	
少林通方正眼(一卷)	釋奉琪
禪源溯流(一卷)	釋有炯
山史畧抄(一卷)	釋有炯
兒菴遺集(三卷)	釋惠藏
日本漂海錄(一卷)	釋賢正
應雲空如大師遺忘錄(一卷)	釋空如
伽山藁(四卷)	釋戒悟
佛說天地八陽神呪經註(一卷)	釋敬和
禪門四辨漫語(一卷)	釋意恂
艸衣詩藁(二卷)	釋意恂
茶神傳(一卷)	釋意恂
東茶頌(一卷)	釋意恂
震默祖師遺蹟攷(二卷)	釋意恂
鐵船小艸(一卷)	釋慧楫
禪學入門(二卷)	金大鉉
述夢瑣言(一卷)	金大鉉
櫟山集(二卷)	釋善影
涵弘堂集(二卷)	釋致能
東師列傳(六卷)	釋覺岸

梵海禪師文集(二卷)	釋覺岸
梵海禪師詩集(二卷)	釋覺岸
優曇林下錄(一卷)	釋洪基
禪門證正錄(一卷)	釋洪基

한국불교전서의 입력은 스캐너를 사용해 책의 내용을 이미지로 저장한 다음, 이 이미지를 OCR 프로그램 ‘글눈2001’을 이용하여 한문을 인식해 내는 방법을 사용하였다. 이렇게 인식된 문서는 많은 탈자 및 오자가 존재하기 때문에 두 차례에 걸쳐 교정 작업을 실시하였다. 다음은 문서를 이미지로 저장하고 인식한 후 초벌 교정을 하는 과정이다.

- ① 한국 불교 전서 원문을 스캔하여 BMP 파일 형식의 이미지를 저장한다.
- ② 저장된 이미지에서 인식할 영역을 선택한 후 인식기를 동작시킨다.
- ③ 인식 과정 중 인식이 되지 않은 부분에 대한 수정작업을 진행한다.
- ④ 인식단계에서 처리하기 힘든 글자는 “?”로 처리한다.
- ⑤ 수정이 완료된 후 외부 편집기(MS WORD 2000)를 실행시켜 수정이 완성된 문서를 붙여 넣는다.
- ⑥ 인식단계에서 “?”로 처리된 문자를 외부 편집기에서 수정한다.
- ⑦ 수정 완료된 문서를 저장한다.
- ⑧ 1차 수정이 완료된 문서에 남아 있는 “?”로 표시되어 처리되지 않은 문자를 원문과 직접 비교해서 독음과 한자를 입력한다.
- ⑨ 2차 수정에서 처리되지 않은 “?”는 추후 이미지 처리를 위해 “&”로 변환한다.
- ⑩ 2차 수정이 완료된 문서를 저장한다.

2차 수정이 완료된 문서를 정확한 웹 서비스 지원을 위해 5차례에 걸쳐 교정 작업을 진행하였다, 교정 작업은 다음과 같다.

- 1) 제1차 교정 : 처음 입력된 것을 틀린 자가 없는지 한글자 한글자 원본과 대조하면서 확인하고, 누락문자로 표시된 & 기호가 정확한 누락문자인지 아니면 이체자 인지를 옥편과 대조해보며, 또한 옥편에 있더라도 반드시 MS WORD 문서에서 인식되는 한문인지를 확인해 본 후 최종 누락문자로 처리하고, 한국불교전서 원문을 대조해가며 세심하게 문장에 정확도를 기하였다.
- 2) 제2차 교정 : 1차 교정본에 제목, 소제목, 쪽수, 단락수, 주석, 들여쓰기 등의 태그 처리를 하면서 다시 한 번 교정함.
- 3) 제3차 교정 : 태그가 다 끝난 후 다시 한 번 전체적으로 오류가 없는지 빠진 누락문자와 태그는 잘 처리되었는지에 대해 교정함
- 4) 제4차 교정 : 태그 처리가 끝나 누락문자를 담당하는 제2팀에서 누락문자 처리가 끝나면 다시 입력팀에서 태그와 문단구성, 누락문자, 이체자 등 전체적으로 오류는 없는지 다시 한번 최종 교정 작업함
- 5) 제5차 교정 : 작업이 모두 끝난 원문이 웹 상에서 제대로 구현되는지를 다시 한 번 살펴본 후 최종 교정 작업을 마친다.

이러한 입력과 5번의 교정 등 여섯 단계를 거쳐 한국불교전서 제9책, 제10책의 전산화가 이루어졌다.

2. 유니코드에서 누락된 문자의 처리

누락 문자란 현재 윈도우즈 운영체제 및 인터넷 환경에서 사용가능한 한자에 포함되지 않는 문자를 뜻한다. 한글 윈도우즈에서 채택하고 있는 KSC-5601 한글 체계상에서 한자는 대략 4,888자 정도 지

원이 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 누락문자(Missing Character)라 칭한다.

이러한 누락 문자가 존재하는 이유는 다음과 같이 볼 수 있다.

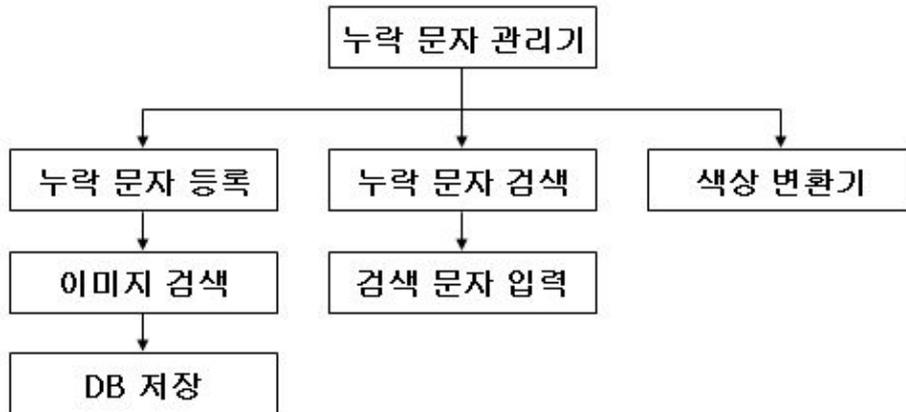
- 고문헌이 집필될 당시의 한자들이 유니코드 내에 포함되어 있지 않은 경우
- 고문헌의 기록 과정에서 오자 입력으로 인한 실제 존재하지 않는 글자인 경우

입력 과정을 거쳐야 하는 한국불교전서의 분량이 방대하기 때문에 가능한 입력 도구의 간소화와 편리화가 필요로 하다. 실제 입력 작업은 스캐너와 OCR을 이용한 방식을 사용하지만 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 지난 과제에서 이미 개발하였고, 본 과제에서는 누락 문자 관리기의 처리 속도와 사용자의 편의성을 향상하였다

4차 사업에 사용하였던 누락 문자 관리기는 일반 사용자가 사용하기에 많은 어려움을 가지고 있었기에 5차 사업에서는 누락 문자 관리기의 사용 편의성을 제공하고 1차에서 4차까지 작성된 누락 문자의 효율적인 관리를 위해 관리기의 개선 작업을 시행하였다.

한국불교전서 전산화 사업이 해를 거듭할수록 관리해야 하는 누락 문자의 수가 점차 증가하여 기존의 누락 문자 관리기를 이용한 누락 문자 관리는 효율성이 떨어지는 한계를 가지게 되었다. 이를 해결하기 위해 개선된 누락 문자 관리기에서 검색의 방법을 용이하게 한다.



[그림 1] 누락 문자 관리기의 구조

[그림 1]과 같이 누락 문자 관리기는 문자의 등록, 검색, 문자의 색상 변환 등 몇 가지 지능별 구조를 갖는다. 문자 등록 기능은 표현되지 않는 문자들의 이미지 추출 및 내부적인 코드 부여, DB에 누락 문자 정보 저장 등의 기능을 가진다. 문자 검색 기능은 등록된 누락 문자의 검색 및 검색된 문자를 원문 상에 삽입하는 기능을 행한다. 끝으로 색상 변환기는 웹페이지에서 사용되는 글자의 색상이 검정색으로 한정되어 있지 않기 때문에 기존의 만들어진 누락 문자 이미지의 색상의 변환이 필요로 하다. 색상 변환기는 사용자가 원하는 색상으로 문자의 이미지를 변환하는 기능을 제공한다.

누락 문자 관리기가 제공하는 검색 기능은 다음과 같다.

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

위의 여러 가지 검색 기능을 제공하기 위해 누락 문자 관리기의 전체 구조 중에서 검색기와 등록기의 기능을 강화하고, 데이터베이스 내의 필드를 추가하여 효과적인 검색이 가능하도록 기능을 확장하였

다. [그림 2]는 확장된 데이터베이스와 데이터베이스에 저장된 내용을 보여주고 있으며 [그림 3]은 누락 문자 관리기의 검색 기능의 인터페이스를 보여준다.

charID	busuID	busuNum	charNum	Pron	CharImage	URL
0548	4-24	4	5	전	<NULL>	
0549	3-32	3	5	도	<NULL>	
0550	4-36	4	5	발	<NULL>	
0601	2-10	2	6	결	<NULL>	
0602	3-33	3	6	장	<NULL>	
0603	3-11	3	6	<NULL>	<NULL>	
0604	3-34	3	6	오	<NULL>	
0605	2-24	2	6	<NULL>	<NULL>	
0606	3-11	3	6	<NULL>	<NULL>	
0607	3-1	3	6	타	<NULL>	
0608	3-7	3	6	<NULL>	<NULL>	
0609	3-10	3	6	<NULL>	<NULL>	
0610	2-4	2	6	기	<NULL>	
0611	2-9	2	6	<NULL>	<NULL>	
0612	3-1	3	6	<NULL>	<NULL>	
0614	4-18	4	6	지	<NULL>	
0615	3-1	3	6	<NULL>	<NULL>	
0616	2-4	2	6	<NULL>	<NULL>	
0617	3-35	3	6	<NULL>	<NULL>	
0618	2-4	2	6	저	<NULL>	
0619	2-6	2	6	미	<NULL>	
0620	2-4	2	6	<NULL>	<NULL>	
0621	4-18	4	6	<NULL>	<NULL>	
0622	2-4	2	6	<NULL>	<NULL>	
0623	3-33	3	6	파	<NULL>	
0624	2-18	2	6	잡	<NULL>	

[그림 2] 개선된 데이터베이스의 데이터

위의 여러 가지 검색 기능을 제공하기 위해 누락 문자 관리기의 전체 구조 중에서 검색기와 등록기의 기능을 강화하고, 데이터베이스 내의 필드를 추가하여 효과적인 검색이 가능하도록 기능을 확장하였다. [그림 2]는 확장된 데이터베이스와 데이터베이스에 저장된 내용을 보여주고 있으며 [그림 3]은 누락 문자 관리기의 검색 기능의 인터페이스를 보여준다.



가. 부수 검색 화면



나. 총획수 검색 화면



다. 음 검색 화면

[그림 3] 누락문자 관리기의 검색 인터페이스

검색 기능을 사용하기 위해서는 데이터베이스에 누락문자를 등록하는 과정이 필요하다. 누락문자를 등록하는 과정은 다음과 같다.

- 모직교 폰트 프로그램을 이용한 누락 문자 이미지 생성
- 부수 선택
- 총 획수 입력
- 만약 누락 문자의 음이 존재하면 음을 입력
- 누락 문자 ID 생성
- 데이터베이스에 누락 문자 저장

누락 문자를 생성하기 위해 가정 먼저 해야 하는 작업은 모직교 폰트 프로그램에서 해당 누락 문자 이미지를 생성하는 작업이다. 생성할 누락 문자 이미지를 찾으면 클립보드에 복사를 하여 임시 저장을 하고 부수, 음, 총획수와 같은 누락문자에 대한 정보를 입력하고 이미지와 정보를 데이터베이스에 저장한다.

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. [그림 4]는 누락문자 등록기의 인터페이스를 보여준다.



[그림 4] 누락문자 등록기 화면

3. 데이터베이스 저장

한국불교전서는 제목, 원문, 주석 등으로 구성되어있고, 각각에 해당되는 내용은 태그로 구별하기 때문에 이를 이용하여 데이터베이스

를 구축할 수가 있다. 이때 원문에 나타나는 한문은 기존 문자 셋으로 표현하는데 한계가 있어서 유니코드로 변환하여 저장한다. 웹에서 검색을 할 때에는 유니코드로 변환된 한자를 다시 원래의 한자로 변환한 형태로 보여준다.

먼저, 원문을 데이터베이스에 저장하기 위해서는 원문을 구별해주는 각 태그들의 유효성을 검증하는 작업이 필요하다. 이러한 유효성 검증 작업을 마친 다음에는 원문으로부터 제목, 원문 내용, 키워드를 추출하여 유니코드로 변환한 후 그 값을 각 테이블에 저장한다. 제목은 tag_jmok_table 테이블에 저장하고, 원문은 edocdata 테이블에 권별로 저장하는데 이번 사업에서는 9책과 10책이 추가로 저장되었다. 또한 키워드는 keyword 테이블에 저장한다. 이러한 한국불교전서의 데이터베이스 구축 단계를 정리하면 다음과 같다.

- ① 태그가 삽입된 원문의 유효성 검증 작업
- ② 키워드 추출 및 저장
- ③ 원문 저장

3.1 태그가 삽입된 원문의 유효성 검증 작업

텍스트 파일로 변환된 원문에 페이지, 제목, 단락, 들여 쓰기, 주석 등을 구별하기 위하여 각각 <PAGE>, <JMOK>, <DAN>, <TAB>, <COMMENT>라는 태그들을 삽입한다. 이러한 태그들은 쌍으로 여는 태그(<...>)와 닫는 태그(</...>)로 구성되어야 하며, 만일 그렇지 않으면 잘못된 데이터가 데이터베이스에 들어갈 수 있으므로 반드시 확인이 필요한 작업이다. 이러한 태그들의 검증 작업은 대략 다음과 같은 순서로 이루어진다.

- ① “*.txt”로 저장된 파일들을 “*.xml”로 확장자 명을 바꾼다.
- ② xml 문서를 웹 브라우저에서 읽어 들인다.
- ③ 웹 브라우저에 아무런 에러 메시지가 안 나타나면 유효한 문서

이고, 에러 메시지가 나타나면 해당되는 내용을 찾아 원문을 수정한다.

- ④ 모든 태그들이 여는 태그와 닫는 태그로 이루어져야만 유효한 문서를 생성할 수 있다.
- ⑤ 최종적으로 이렇게 생성된 유효한 문서를 데이터베이스 구축에 사용한다.

3.2 키워드 추출 및 저장

키워드 추출 및 저장 단계에서 이뤄지는 작업은 원문의 저장 단계에서와 동시에 이루어지는데, 원문 내에서 키워드로 지정된 단어를 찾아 그 위치와 단어를 keyword_index 테이블에 저장하는 작업을 한다. 키워드에 관한 정보는 ekeyword, hkeyword, stroke 테이블에 저장되어 있는데, 각각 키워드의 유니코드 값, 한글 값, 획수 값을 저장하고 있다. 이러한 키워드 추출 및 저장 작업은 대략 다음과 같은 순서로 이루어진다.

- ① ekeyword 테이블로부터 키워드의 목록을 해쉬(hash) 테이블 자료구조 형태로 구축한다.
- ② 원문에서 한 라인씩 입력으로 받은 문자열 내에서 ①의 해쉬 테이블을 이용하여 키워드들을 찾는다.
- ③ 키워드의 해당하는 단어들을 keyword_index 테이블에 저장한다.
- ④ 중복된 키워드를 제거하여 권마다 유일한 키워드 목록만을 따로 keyword 테이블에 저장한다.

3.3 원문 저장

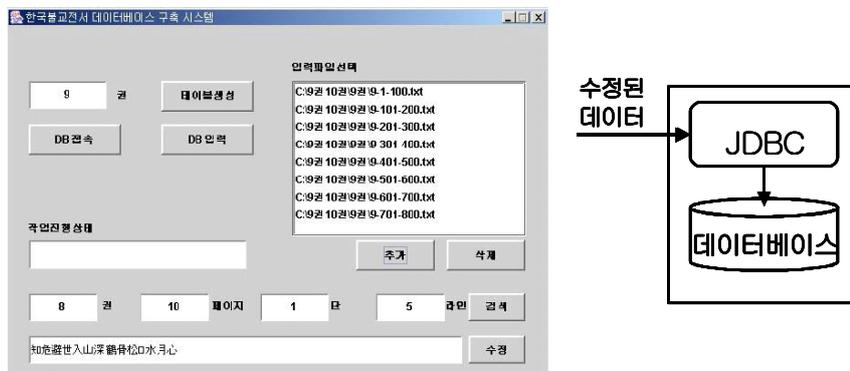
원문 저장은 유니코드 편집기에서 작성된 유니코드 원문을 그대로 테이블에 저장하는 단계이다. 원문 파일을 라인(line)별로 읽어 원문의 내용을 저장하면서 페이지 태그와 단 태그를 검사하여 페이지 당

라인 수와 단 번호 등의 부가 정보를 생성한다. 이러한 부가정보는 원문에 대한 인덱스 역할을 한다. 원문 저장에 사용되는 edocdata 테이블의 속성은 다음과 같다.

테이블 명: edocdata		
칼럼명	데이터형(길이)	비고
nlinenum	integer	라인 수
sdocdata	nchar(800)	유니코드 원문저장
npagemum	integer	페이지 번호
npageline	integer	페이지에서의 라인 번호
ndannum	integer	단 번호

3.4 데이터베이스 원문 저장 시스템 구현

[그림 5]는 데이터베이스 저장 시스템의 구조도를 나타낸 그림이다. 즉, 원문 저장 프로그램을 이용하여 원문 데이터를 입력하면 이를 데이터베이스에 저장한다. 이때, 프로그램과 데이터베이스간의 연동을 위하여 JDBC를 사용하고, DBMS로는 마이크로소프트사의 SQL 7.0을 사용하였다. 다음에서 데이터베이스 원문 저장 프로그램의 주요 기능은 원문 저장 기능, 원문 검색 기능, 원문 수정 기능이 있다.



원문 저장 프로그램

[그림 5] 데이터베이스 원문 저장 시스템 구조도

원문 저장 기능은 원문을 저장하는 기능이고 원문 검색 기능은 원문 저장 프로그램의 입력된 데이터에 대한 검색 기능이며 원문에서 검색하고자 하는 내용을 찾을 수가 있을 뿐만 아니라 데이터의 오류 부분을 쉽게 찾아낼 수 있다.

마지막으로 원문 수정 기능은 원문 검색 기능을 이용하여 오류가 있는 원문 부분을 찾아낸 후에 수정을 할 수 있어 원문의 양이 많아져서 원문 전체를 전부 입력할 때 생기는 시간과 노력을 줄이는 효과를 보인다.

4. 웹 검색 인터페이스의 구현

한국불교전서가 제목, 원문, 주석 등으로 구성되어있기 때문에 검색 결과 화면도 경전과 동일한 구성으로 제공 하였다. 사용자가 쉽고 편리한 방법으로 검색 할 수 있도록 키워드 검색, 페이지 검색, 제목 검색, 그리고 획수 검색에 이르는 다양한 검색 서비스를 제공하였다. 또한 원문과 동일한 형태로 들여쓰기 기능을 제공함으로써 사용자에게 학술적인 참고 자료로서 가치가 있도록 하였다.

웹 검색 인터페이스 측면에서 수행한 5차 사업의 내용은 몇 가지 사용자의 요구사항을 추가하였다. 추가한 요구사항은 원문의 형태에 가깝게 들여쓰기 기능, 선택한 항목강조 기능, 소스코드 보안기, 그리고 틀 고정기능이다. 웹 검색 인터페이스의 주요 기능과 제5차 사업에서 수행한 내용은 본문을 통해서 살펴보도록 한다.

4.1 웹 검색시스템의 주요 기능

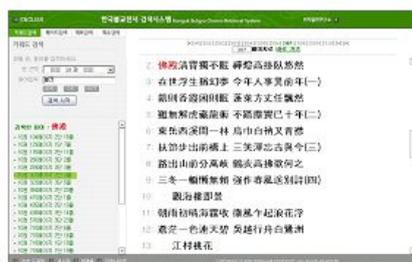
한국불교전서 웹 검색 인터페이스는 사용자에게 보다 편리한 검색을 위해서 여러 가지 검색 방법을 제공하고 있다. 주요 검색 방법으로는 키워드 검색, 페이지 검색, 제목 검색, 그리고 획수 검색으로 구성되어 있다. [그림 6]은 웹 검색시스템의 주요 기능의 인터페이스를

보여준다.

키워드 검색은 경전의 키워드를 이용해서 검색을 하며, 다양한 검색 조건을 처리할 수 있는 기능을 제공한다. 페이지 검색은 경전을 검색할 때 찾으려는 페이지를 직접 입력해서 검색하는 방법이며 제목 검색은 경전의 각 ‘권’에 포함되어 있는 제목을 이용하여 검색하는 방법이다. 마지막으로 획수 검색은 한자의 획수를 이용해서 검색하는 방법이다.

4.2 추가 기능

본 검색시스템은 사용자에게 유용하고 편리한 기능을 제공하고 완성도 높은 웹 검색시스템을 위하여 새로운 사용자 요구사항을 추가하였다. 그 내용은 들여쓰기 기능, 선택한 제목강조 기능, 소스코드 보안기능, 그리고 틀 고정 기능 등이다. 각 기능에 대한 내용을 좀더 상세하게 설명하면 다음과 같다.



가. 키워드 검색 화면



나. 페이지 검색 화면



다. 제목 검색



라. 획수 검색 화면

[그림 6] 웹 검색 시스템의 주요 기능 인터페이스

(1) 들여쓰기 기능

한국불교전서에서 경전의 첫째 칸부터 나오는 내용은 경전 본문을 의미하고, 둘째 칸부터 나오는 내용은 해설을 의미한다. 들여쓰기 기능은 원문과 동일한 형태를 제공함으로써 사용자에게 학술적인 참고 자료로서 가치가 있도록 하였다.

(2) 선택한 항목강조 기능

이 기능은 사용자가 키워드 검색, 페이지 검색, 제목 검색, 그리고 획수 검색을 하고 나면 그 검색 항목이 좌측 틀에 나타난다. 사용자가 원하는 항목을 선택하면 강조의 효과로 선택한 항목에 음영이 나타나도록 하였다. 이 기능을 제공함으로써 본 검색시스템의 사용자는 현재 사용자가 선택한 항목이 무엇인지를 항상 명확하게 보여 줄 수 있다.

(3) 소스코드 보안 기능

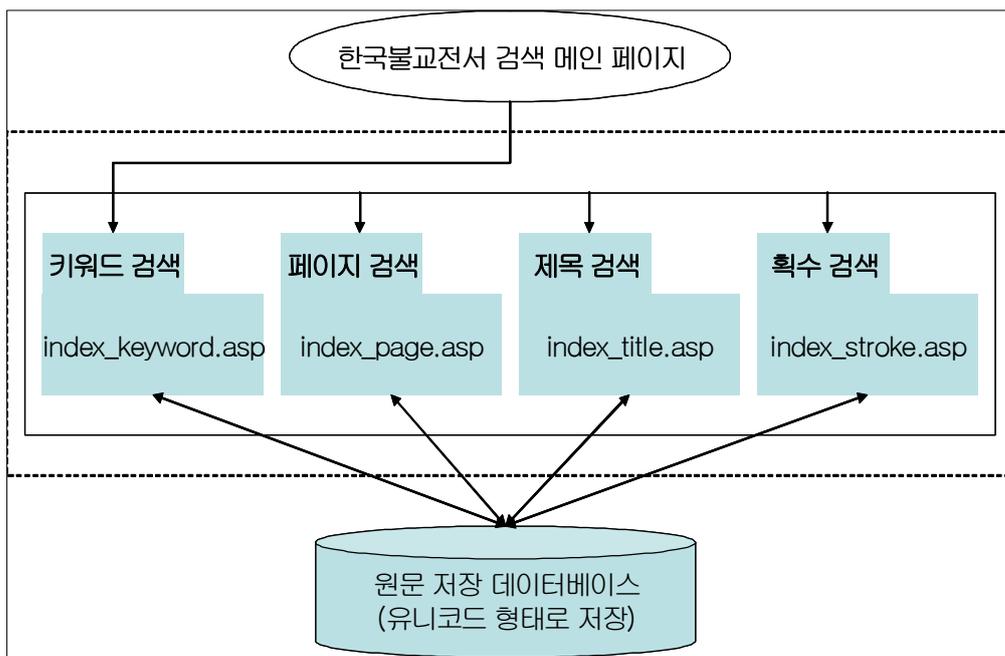
소스코드란 프로그램이 어떻게 작동할지를 결정하는 것으로서 소프트웨어 산업에 있어서 중요한 자산이자 기술에 해당한다고 할 수 있다. 본 검색시스템에서도 소스코드가 중요한 기술 자산에 해당한다고 인식을 하였기 때문에 소스코드에 대한 보안기능을 제공하였다.

(4) 틀 고정 기능

이 기능은 좌측 틀의 검색부분을 고정함으로써 검색한 후 이동막대를 움직여도 검색부분은 변함없이 그 위치에 고정이 되도록 하였다. 이렇게 함으로써 사용자는 고정된 검색부분을 이용해서 편리하게 검색을 할 수 있다.

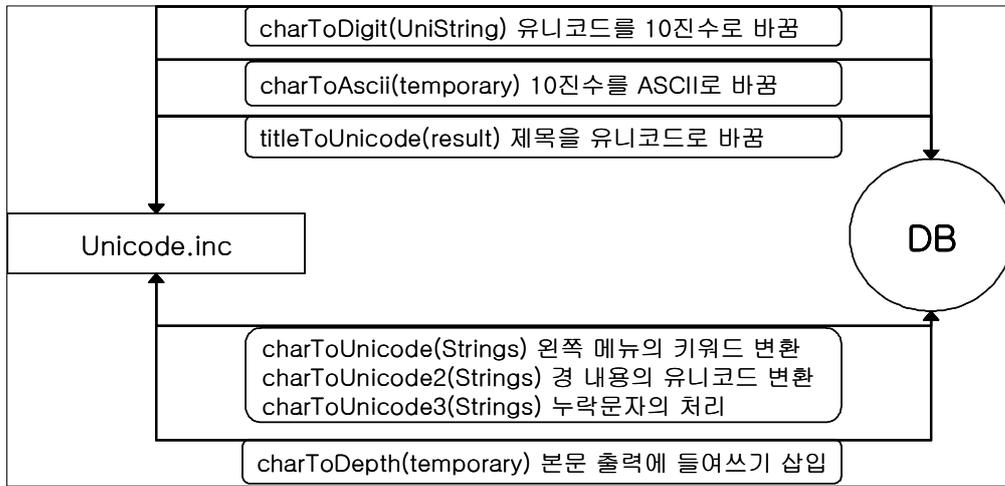
4.4 웹 검색 인터페이스의 구현

[그림 7]은 본 검색시스템의 전체 구성도를 나타낸 그림이다. 사용자가 메인화면에서 검색시작 단추를 누르면 기본적으로 키워드 검색 페이지로 이동한다. 페이지 검색, 제목 검색, 그리고 획수 검색으로 이동하기를 원할 때는 해당 탭을 누르면 된다. 사용자가 원하는 검색 결과를 얻는 과정은 해당 검색 페이지에서 검색을 요청하면, 본 검색 시스템이 사용자의 검색 요청을 질의문으로 변경한다. 그 다음 원문 저장 데이터베이스에 질의하며, 해당 결과를 사용자에게 보여준다.



[그림 7] 웹 검색 인터페이스의 전체 구성도

또한 데이터베이스에 저장할 때 경전의 내용은 유니코드 형태로 변환해서 저장한다. 그러므로 유니코드와 텍스트간의 변환 기능이 필요하고, [그림 8]은 유니코드와 일반 텍스트간 변환을 담당하는 함수와 변환 과정을 나타낸 그림이다.



[그림 8] 유니코드와 텍스트간 변환 함수

III. 결론 및 향후 연구과제

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 한국불교전서의 일부를 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. 따라서 본 연구를 수행하면 한국불교전서 제9책과 제10책을 전산화 하여 이를 연구하는 연구자들이나 열람을 원하는 사람들에게 도움이 될 뿐만 아니라 우리의 귀중한 문화유산을 전 세계에 널리 알릴 수 있다.

한국불교전적은 우리나라에 불교가 전래된 이래 삼국시대부터 우

리나라의 선조들이 남긴 옛 문헌들을 발굴 수집하여 출간한 한국불교전서와 해인사 고려대장경을 동국대학교의 역경원에서 한글로 번역한 한글 대장경을 위시한 한국의 불교 문헌을 총칭한다. 이러한 불교 경전에 대한 활발한 편찬 사업에 비해 전산화 작업이 현재는 미비한 상태이나 전 세계적으로 연구되고 있고, 이러한 디지털 경전에 대한 정보 교환을 위해 국제 회의인 EBTI가 개최되고 있다. 불교정보화와 관한 유일한 국제회의인 EBTI는 인터넷에 제공되고 있는 다양한 불교정보를 서로 자유롭게 이용할 수 있는 호환성을 키우는데 그 중요성이 있다. 이 회의를 통해 각국에서 진행하는 불교정보화에 대해 현재까지 진행된 상황 등 여러 가지 기술에 대한 의견을 교환한다.

이렇게 EBTI에서 논의되고 있는 기술들은 본 연구를 수행하는데도 필요한 기술로 가장 먼저 한자를 컴퓨터에 입력할 수 있는 입력 방법 및 유니코드 상의 없는 고문헌 상의 문자를 처리할 수 있는 시스템의 개발이 필요하다. 이러한 기술을 개발하기 위해서는 유니코드 등 세계 여러 나라의 각국 언어에 대한 코드 체계 및 방대한 량의 한문 폰트의 확보가 시급하다. 따라서 본 연구소에서는 일본 모직교의 9만 여자의 폰트와 이에 대한 이미지를 이용하여 문헌 중 유니코드 상의 없는 문자를 처리할 수 있는 문자 관리 시스템을 개발하였다. 개발된 문자 관리 시스템을 통해 한국불교전서 제9책과 제10책의 내용을 입력하고, 입력된 내용들을 3번씩 교정 작업을 하여 원문과 다른 글자가 입력되거나 원문에 있는 내용이 빠진 경우들을 없애고 최대한 원문에 가깝게 컴퓨터에 입력하였다.

그리고 본 연구를 수행하는데 두 번째 필요한 기술은 입력된 한국불교전적 원문 내용들을 의미있는 단위로 분할하여 데이터베이스에 저장하는 것이다 또한 저장된 데이터베이스에서 사용자가 질의를 하면 그 질의에 대한 효율적으로 검색할 수 있는 검색 기술 및 한국불교전적 색인파일 작성 기술이 필요하다. 따라서 크게 다음의 4단계로 데이터베이스 구축을 하였다. 가장 먼저 원문에서 키워드를 추출하여

테이블로 저장하는 동시에 인덱스를 구축하기 위한 파일을 생성하는 단계, 다음은 원문 저장을 XML 태그들을 유지하고 원문의 라인을 유지하면서 저장하는 단계, 그리고 인덱스 구축 및 문서 구조 추출 순서로 이루어진다.

마지막으로 필요한 기술은 데이터베이스에 저장되어 있는 내용들을 검색하기 위하여 전 세계에서 사용하고 있는 인터넷의 웹을 통해 검색할 수 있는 웹 인터페이스와 인터넷이 되지 않는 환경에서도 사용자가 질의를 입력하면 이들을 검색할 수 있는 CD-ROM을 통한 검색 방법이 필요하다. 이에 본 연구에서는 위에서 언급한 3가지 기술들과 사용 방법들을 개발하였다.

본 연구에서 개발된 한국 불교전서 제9책과, 제10책에 대해 웹을 통해 검색하고자 한다면 다음의 URL을 이용하면 된다. URL은 <http://ebti.dongguk.ac.kr/> 이다. 향후 연구 과제는 한국 불교전적 전책을 다 빠른 시일 내에 전산화하는 일이 필요하다. 더불어 본 연구에서 개발된 유니코드에서 누락된 문자 처리 시스템에서 보다 효율적이고 빠른 체계적인 문자의 관리를 위해 기능을 수정 보완해야 한다. 그리고 데이터베이스에 저장된 내용을 검색할 때 원문 전체에 대한 전문 검색 방법도 가능하도록 해야 하겠다.

참고 문헌

- [1] Aming Tu, “중국 전자 불전 협회(CBETA)의 전자 『大正新脩大藏經』,” ’01 동국대학교 개교 95주년 기념 세계전자불전학회 학술대회, 2001.
- [2] Fred Coulson, “전기(傳記)-저서(著書) 목록 검색 데이터베이스로 링크된 텍스트 이미지를 위한 TBRC와 그 모델들,” ’01 동국대학교 개교 95주년 기념 세계전자불전학회 학술대회, 2001.
- [3] John Lehman, “탈자(脫字) 문제 처리를 위한 프로젝트,” ’01 동국대학교 개교 95주년 기념 세계전자불전학회 학술대회, 2001.
- [4] Robert Chilton, “아시아 고전 입력 프로젝트 (ACIP): 과거, 현재

그리고 미래,” ’01 동국대학교 개교 95주년 기념 세계전자불전학회 학술대회, 2001.

- [5] Eric Johnson, The Text Encoding Initiative, Text Technology, 1995.
- [6] ISO 8879, Standard Generalized Markup Language, 2nd Edition, 1986.
- [7] ISO/IEC 10646-1, “Information Technology – Universal Multiple-Octet Coded Character Set(UCS) – Part I : Architecture and Basic Multilingual Plane,” 1993.
- [8] The Unicode Consortium, The Unicode Standard, Version 2.0, Addison Wesley, 1996.
- [9] The Unicode Standard, Microsoft Developer’s Network, 1997.
- [10] Unicode Enabling, Microsoft Developer’s Network, 1997.
- [11] Unicode Support in Win32, Microsoft Developer’s Network, 1997.
- [12] CJK Codes–Unicode/ISO–10646 Unicied “Ideographs,”
<http://www.mit.edu:8001/afs/athena.mit...r/a/k/akbar/www/Unicode-ideographs.html>.
- [13] Christian Wittern, “Chinese character codes: an update,”
<http://www.iiinet.or.jp/iriz/irizhtml/multling/codes/htm>.
- [14] EditTime, <http://www.timelux.lu>, TimeLUX.
- [15] How to View Chinese/Japanese/Korean HTML with Netscape Communication on US version of Windows 95 or NT,
[hpeople.netscape.com/ftang/communicatorfont.html](http://people.netscape.com/ftang/communicatorfont.html).
- [16] “Installing Bitstreaan Cyberbit Version 1.1,”
<http://www.bitstreaan.com/cyberbit.html>.
- [17] “Notes on CJK Character Codes and Encodings,”
<http://www.ifcss.org/ftp-pub/software/info/cjk-codes>.
- [18] Panorama, <http://www.softquad.com>, Softquad.

- [19] Public Unicode Font,
<ftp://www.ifcss.org/ftp-pub/software/fonts /unicode>.
- [20] True Type and Unicode,
<http://truetype.demon.co.uk:80/unicode.htm>.
- [21] Urs App, “A Look at the Korean Tripitaka Input Project,”
<http://www.ijnet.or.jp/iriz/irizhtml/ebit/samsung.htm>.
- [22] Urs App, “Guidlines for the Creation of Large Chinese Text Databases,”
<http://www.ijnet.or.jp/iriz/irizhtml/maketext/guideline.html>.
- [23] Urs App, “The Importance of Markup,”
<http://www.ijnet.or.jp/iriz/irizhtml/maketext/foguang.html>.
- [24] 대장경학술용어연구회, “대정신수 대장경소인,” 제1권, 대장경학술용어연구회, 1975.
- [25] 송석구, “전자불전과 미래불교의 향방,” '01 동국대학교 개교 95주년 기념 세계전자불전학회 학술대회, 2001.
- [26] 이금석, “한국불교전서 전산화에서의 누락문자관리,” '00 동국대학교 전자불전연구소 제2회 세미나, 2000.
- [27] 한태식, “불교학 연구에 있어서 한국불교전서의 위상,” '00 동국대학교 전자불전연구소 제2회 세미나, 2000.
- [28] 허인섭, “Report on the Digital Tripitaka Koreana 2001,” '01 동국대학교 개교 95주년 기념 세계전자불전학회 학술대회, 2001.
- [29] 현득창, 임광택, 이수연, “SGML 기본 파서를 이용한 SGML문서 편집기의 구현,” 한국정보과학회, 정보과학회 논문지, Vol 25, No. 1, 1998.
- [30] 홍영식, “한국 불교전서 데이터베이스에서 누락 문자 검색,” '01 동국대학교 개교 95주년 기념 세계전자불전학회 학술대회, 2001.
- [31] 김숙자, SGML의 모든 것, 성안당, 1997.
- [32] 동국대학교 출판부 발행, 한국불교전서 제9-10권 조선시대편, 1979.

- [33] 장희창, 현득창, 이수연, SGML 가이드, 사이버출판사, 1997.
- [34] 한국불교신문, 시방시계 1월 27일자, 현대불교신문사, 1999.
- [35] 황기태 역, 어드밴스 윈도우 NT, 도서출판 대림, 1995.
- [36] 강석진, “팔만사천대장경 전산화를 위한 제언, 한자위주 문헌의 워드프로세서 데이터베이스, 탁상출판 시스템 개발을 위해,”
<http://members.iWorld.net/hederein/menu22/Kang.html>.
- [37] 김응철, “고려장경 및 한자정보전산화에 관련한 문제제기,”
<http://members.iWorld.net/hederein/menu22/Kim.html>.
- [38] 노용균, “불전 전산화와 SGML,”
<http://members.iWorld.net/hederein/menu22/Dogam42.html>.
- [39] 심재룡, “정보화 사회와 불교 전산화,”
<http://members.iWorld.net/hederein/menu22/Dogam32.html>.
- [40] 이규갑, “고려대장경 전산화에 있어서 이체자의 처리 문제,”
<http://members.iWorld.net/hederein/menu22/Yi.html>.
- [41] 인터넷으로 만나는 불교,
<http://members.iWorld.net/hederein/menu23/Pogyu121.html>.
- [42] 정주원, “ISO/IEC-10646 Universal Multiple-Object Coded Character Set (UCS)에 대해서,”
<http://simac.kaist.ac.kr/~jwjung/seminar/hangul-i18n/iso10646.html>.
- [43] 정주원, “한글 코드에 대하여,”
<http://simac.kaist.ac.kr/~jwjung/seminar/hangul-i18n/ko-code.html>.
- [44] 종림스님, “팔만대장경 전산화 추진경과와 이후 계획,”
<http://members.iWorld.net/hederein/menu22/>
- [45] 혜묵스님, “세계의 불교자료 전산화 계획과 고려대장경 전산화를 위한 몇가지 문제들,”

<http://members.iWorld.net/hederein/menu22/Hye.html>.

키워드(Keyword)

한국불교전서, 한국불교전서 검색 시스템, 한국불교전서 전산화,
유니코드, 누락문자

Korea Bulgyo Chonso, Korea Bulgyo Chonso Retrieval
System, Korea Bulgyo Chonso Digitalization, Unicode, Missing
Character