학위논문의 원문DB 구축의 실제

-XLX 포맷을 중심으로-

서울시립대학교 중앙도서관
허 경 희

〈목 차〉

Ⅰ. 서 론

전통적인 도서관은 인쇄매체 자료를 수집·정리하고, 축적하여 이들을 이용자에게 제공하여 왔으나 컴퓨터와 통신의 발달로 현대의 도서관은 대부분 자동화 시스템을 도입하여 수작업으로 하던 기존의 도서관 업무를 전산처리하고 있다. 그러나 도서관 자동화 시스템은 원 문헌에 대한 논리적 및 물리적 구조가 무시된 메타데이터만을 제공하는데 그치고 있다. 이러한 점을 보완하기 위하여 근래에는 대학도서관에서 보다 높은 서비스를 하기 위해서 전자도서관 사업을 진행중이거나 진행할 준비를 하고 있다.

전자도서관에서는 궁극적으로 도서관 이용자가 멀티미디어 PC 1 대에서 인터넷과

같은 통신망을 통해 전 세계 각처에 제각기 구축되어진 다양한 정보매체를 검색하여, 원하는 원문정보를 즉각적으로 제공받을 수 있다.

본교 도서관의 전산화 현황은 TG-VintageLAS의 문헌정보시스템을 통하여 수서, 편목, 대출·반납, 연속간행물관리, 기사색인 관리 등의 업무를 자동화 하였고, 서지정보를 검색터미널과 홈페이지를 통하여 제공하고 있으며, 메타데이터인 서지정보 뿐 아니라 원문정보도 제공하고자, 종합정보화 1차년도 사업으로 본교 석·박사 학위논문을 대상자료로 하여 1983년도-1999년도 논문을 원문DB 구축하여 Web기반의 검색서비스를 실시하고 있다.

본고에서는 본교 도서관에서 학위논문 원문 DB를 구축하는데 사용한 원문정보시스템에 대해서 기술하고자 하며, 전자포맷인 XLX를 중심으로 기술하고자 한다.

2. 인쇄문헌의 디지털화 방법

2. 1. 텍스트형태의 디지털 방법

텍스트 형태의 디지털 방법은 본문에 대한 접근이 전제되므로 검색의 확장성을 높일 수 있으며 향후 디지털도서관을 구축하는 방향이다. 텍스트 포맷에는 HTML, SGML, XML, PDF, LaTex, XLX 등이 있다.

HTML(Hyper Text Markup Language)은 SGML을 기반으로 한 언어의 일종으로 1990년대 이후 인터넷의 문서를 기술하는 대표적인 언어로 발달해 왔다. 그러나 HTML은 특별한 데이터 타입이 사용되지 않고 단순한 텍스트이기 때문에 인쇄문헌이 지닌 다양한 특성을 나타낼 수 없고, 또한 구조적으로 문서를 표현하지 못해 전자문서를 구현하는데 많은 제약이 따른다.

SGML(Standard Generalized Markup Language)은 ISO (International Standard Organization) 8879(1986)에서 정의한 표준 언어이고 문서의 논리구조를 일반적으로 기술하는데 쓰이는 마크업 언어를 정의하기 위한 언어로써 이 문법을 어떻게 적용하느냐에 따라 동일한 문서도 다르게 표현될 수 있다. 마크업이란 편집자나 인쇄 디자이너가 색인, 텍스트 및 그 외의 부분에 대해서 서체나 사이즈 등 체제에 관한 지정을 원고 상에 손으로 기입하는 방식이다. SGML은 문서 데이터베이스의 작성이나 유지, 보수에 유효할 뿐만 아니라, 교환 유통에도 적당하다. 개개의 시스템에 독립되어 있을 뿐만 아니라 구조가 엄격히 정의되어 있어, 개개

의 시스템을 위하여 변환할 수 있기 때문이다. 나아가 SGML에서 만들어진 문서 데이터베이스는 전자도서관의 시스템의 자원으로도 사용된다. 그러나 SGML은 적용시 아직도 개발해야 될 기능이 많아서 이의 완벽한 처리는 기간이 필요하다.

XML(EXtensible Markup Language)은 SGML에서의 복잡성을 제거하고 HTML에서의 고정된 태그를 벗어나 사용자가 문서구조를 정의하여 사용할 수 있는 것으로 SGML과 HTML의 단점을 상호 보완한 문서기술언어이다. 현재 WWW 컨소시움에서 웹의 유용성을 높이기 위해 공식적으로 권하고 있다. HTML이 웹 페이지의 내용(문장과 그림의 배열, 관계)을 기술하는 데 비해 XML은 데이터가 기술되는 용어로 내용을 기술한다. XML은 아직은 국내에서 널리 사용되지 않고 있다.

PDF(Portable Document Format)는 Windows, Mac, UNIX등 다양한 운영체제에서 호환되며, 압축기능을 가지고 있어서 인터넷 등에서 빠르게 파일을 전송할 수 있다. 원본과 동일한 모습을 보이고 서체 및 페이지 포맷정보를 내장하고 있으며, 이미지를 포함하고 있기 때문에 텍스트 파일과 이미지 파일 모두를 PDF파일로 변환할 수 있다. 북마크, 하이퍼링크, 내용검색 등이 가능하고 뷰어도 인터넷상에서 쉽게 받을 수 있다. PDF(Portable Document Format)는 미국의 Adobe사에서 개발한 전자문서 포맷으로 해외 특히 영미 권에서는 표준에 준할 만큼 통용되고 있지만, 국내에서는 한글 지원 제품이 영문판보다 늦게 나온 데다 워드프로세서 아래ᄋᆞ한글에서 지원되지 않고 가격조차 비싸 확산이 늦었다. 그러나 최근에 한글판 Acrobat과 HWP를 PDF로 변환시키는 전용 변환툴을 아래ᄋᆞ한글 워드프로세서와 한 데 묶어 10만원 미만의 저렴한 가격에 출시함으로써 기술과 가격 두 가지 문제가 동시에 해결되어 단 시간내 시장을 장악해 나갈 것으로 예상된다.

LaTeX는 이공계분야 사람들이 이용하는 수학 공식을 잘 표현해 낼 수있다. 대부분의 LaTeX시스템은 이용자가 직접 아스키 파일을 LaTeX 코드로 마크업 한 후에 디스플레이와 인쇄를 위해 Postscript와 같은 포맷으로 변환시켜야 한다. 먼저 DVI(DeVice Independent)포맷으로 만들기 위해 해당 파일을 LaTeX포맷기로 처리하고, 그 다음에 Postscript로 변환시킨다. LaTex는 Tex라고 불리는 보다 낮은 수준의 마크업 언어로 만들어진다.

XLX(PDL XL eXtension)는 PDF와 유사한 국내에서 개발된 포맷이다. 파일의 포맷을 PCL6 (PDL XL)으로 변환하여 XLX파일형식으로 통합 관리하는 것이다. XLX는 윈도 어플리케이션 상의 내용을 프린터에서 파일로 출력한 결과이다. XLX문서는 데이터 코드 값을 가지고 있으므로 페이지 단위까지 검색이 가능하며 전

송 속도가 빠르다. 워드프로세서 파일을 XLX형식으로 변환하면 파일크기가 워드프로세서의 종류에 따라 다소 차이가 있지만 대체로 줄어든다.

텍스트 형태의 DB의 단점은 화면상의 디스플레이 문제와 텍스트의 편집과정이 포함되기 때문에 인쇄문헌의 모습을 완벽하게 재현하지 못할 수 있다. 또한 사진과 그래픽자료는 나타내기가 곤란하다. 사진과 그래픽 자료는 보통 이미지를 스캔하여 텍스트 형태의 자료 중에 삽입시키는 방법이 사용된다. 장점은 텍스트 형태의 DB는 화면에 디스플레이 되거나 프린터에 출력될 경우 미려하고, 파일 저장용량이 작으며, 이미지 형태에 비해 전송속도가 빠르다. 그리고 내용전체를 대상으로 검색이 가능하다는 점이 가장 큰 장점이다.

2. 2. 이미지 형태의 디지털 방법

이미지 포맷의 종류에는 GIF(Graphic Interchange Format), JPEG(Joint Photographic Experts Group), TIFF(Tagged Image File Format), PDF (Portable Document Format)등이 있는데, 문서자료에 대한 포맷으로는 주로 TIFF 나 PDF가 사용된다.

GIF(Graphic Interchange Format)는 미국의 CompuServe사가 컬러그래픽 정보를 축적하고 전송서비스하기 위하여 고안한 것으로 산업계 표준으로 흔히 사용되고 있다. 그리고 현재 웹상의 문서표현으로 이미지 처리는 대부분 GIF를 사용한다.

JPEG(Joint Photographic Experts Group)은 원래 표준을 만들었던 위원회의 이름을 나타내며, JPEG은 원문을 이미지화 했을 때 파일의 크기가 작으며 사람의 눈으로 보기에는 원문과 거의 차이가 없어 사진이나 그림 등이 포함된 일반문헌의 이미지화에 효과적으로 사용할 수 있다. JPEG제정의 주요한 목표는 고도화된 최신의 압축 알고리즘을 채택하여 높은 압축률을 실현하고 동시에 원문의 높은 재현성을 유지하고자 하였으며, 압축기술을 실무에 적용하고자 할 경우에 구현하기 쉽게 논리를 정립하고 소프트웨어로 처리하더라도 충분한 성능을 낼 수 있게 하는데 있다. JPEG은 문헌을 스캐닝하면 페이지 수만큼의 파일이 생성되어 페이지를 각각의 파일로 저장하여야 하고, 많은 파일을 관리해야 하는 단점이 있지만 웹브라우저 상에서 직접 볼 수 있는 장점이 있다.

TIFF(Tagged Image File Format)는 미국의 Aldus사 제품인 'FreeHand"에서 구현하여 이미지 데이터의 정보 교환용으로 널리 사용되고 있는 형태이다. TIFF형식에는 JPEG처럼 한 페이지를 각각 파일로 저장하는 Single TIFF와 전체 페이지를 하나의 파일에 저장하는 Multi TIFF가 있다. Multi TIFF는 이미지 정보에 목차, Page

Matching정보 등의 표시가 가능하며, 이미지 원문에서 처리하기가 불가능한 목차정보를 텍스트로 처리할 수 있다. 그러나 칼라문서는 표현해야 할 정보가 많아서 TIFF방식은 원문의 정보를 누락시켜 양질의 이미지 데이터를 처리하는데는 부적합하다.

근래에는 PDF(Portable Document Format)가 텍스트 파일뿐만 아니라 이미지 파일을 수용하고, 마크업 등의 편집이 가능하기 때문에 널리 사용되고 있다.

이미지 형태의 DB는 텍스트 형태보다 파일의 크기가 크고, 내용전체를 대상으로 하는 검색을 할 수 없는 단점이 있지만, 원문의 모습을 그대로 전하기 때문에 인용면에서 장점이 있다.

3. 원문정보시스템

원문정보시스템은 삼보정보시스템과 오름컴퓨터가 공동으로 개발한 시스템으로 현재 전국의 대학 및 전문, 특수도서관에서 많이 사용하고 있는 도서관자동화 프로그램인 TG-VintageLAS의 편목, 검색 모듈과 연동하여 원문이미지자료를 입력, 편집, 저장 검색할 수 있도록 설계된 TG-Imagebase 와 포항공과대학교가 개발하고 고원시스템이 공급하고 있는 도서관자동화 프로그램인 LINNET과 연동하여 원문을 입력하고 검색할 수 있는 원문 이미지시스템 등이 있다. 본 도서관에서는 LG-EDS의 원문정보시스템을 이용하여 원문DB를 구축하였으며, 원문 Viewer, 원문의 구조정보 관리기능, 원문 데이터의 Conversion 기능 등을 위해서는 JetDocument 제품군을 이용했다 . LG-EDS의 원문정보시스템은 원문 관리 시스템, 원문 웹 검색시스템, Z39.50시스템으로 구성되어 있다.

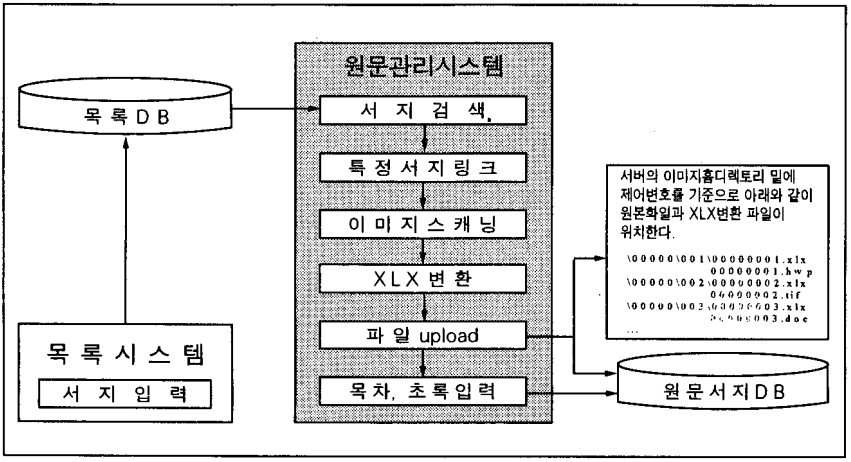
3.1. 원문정보시스템 기자재 현황

본 도서관의 원문정보시스템의 기자재 현황은 다음과 같다.

- 원문 Server - 1 Set
  - 품명 : SUN E450
  - Memory : 512 MB
  - CPU : 300MHz/2MB x 2
  - 원문관리/검색서비스 S/W 등
- Scanner : FUSITU 3097DG

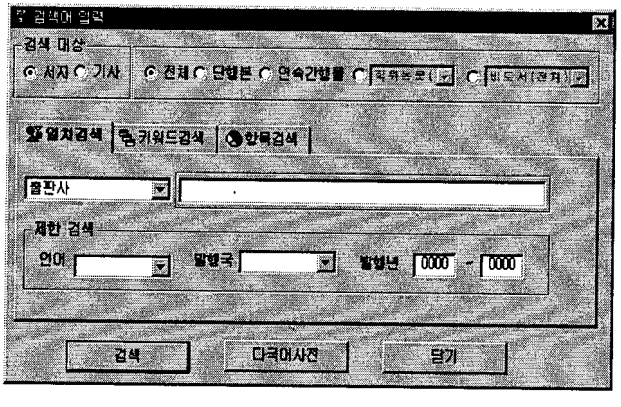
3. 2. 원문 관리 시스템

1) 구성도



2) 서지검색

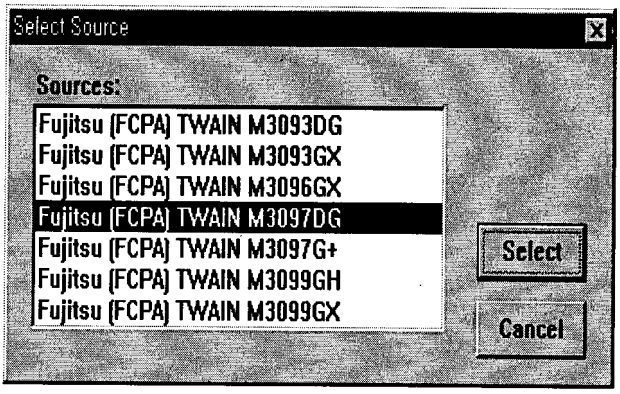
원문 구축 대상이 되는 데이터의 서지사항을 검색하며, 검색방법으로는 일치 검색, 키워드 검색, 항목 검색 등이 있으며. 언어, 발행국, 발행년으로 제한하여 검색을 할 수도 있다.



3) 이미지 스캐닝

인쇄문헌을 스캐닝 하여 TIFF파일을 생성한다.

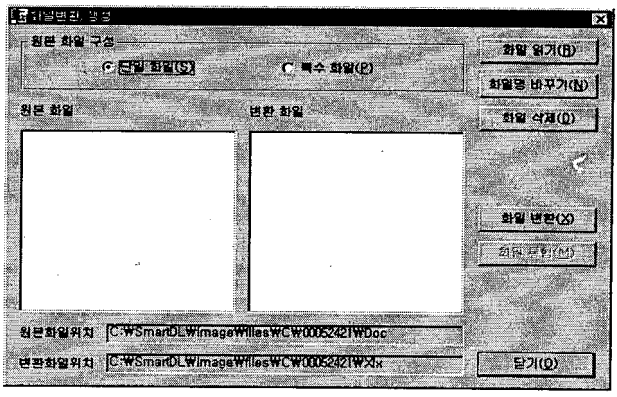
스캐닝시 적절한 해상도를 지정하여야 하며 이때 지나치게 해상도를 높이면 파일의 크기가 커지고 스캐닝 속도가 늦어진다. 스캐닝은 페이지별로 이루어진다.

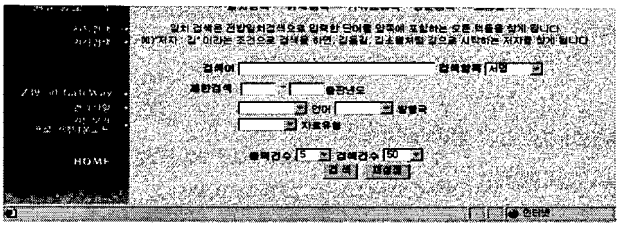


4) 파일변환

인쇄자료는 Multi TIFF형식으로 스캐닝을 한 후 XLX 파일로 변환하고, 디스켓이나 CD자료는 워드 파일(doc, hwp등)을 읽어 들인 후 XLX 파일로 변환한다.

default로 A드라이브를 읽어들인후 XLX파일로 변환하며, 파일 변환시 주의할 점은 원본파일이 작성된 응용프로그램이 실행되는 경우가 있는데 이때 작업절차를 유지하기 위하여 바이러스 체크와 같은 기타의 응용프로그램은 모두 종료한다.





5) 구조정보 작성하기

북마크화면을 열고 목차의 내용에 따라 특정 워크폼을 선택한 후 목차를 작성한다. 그리고 특정 북마크의 내용을 수정하여 목차를 작성한 다음 원문의 지정된 페이지의 내용과 일치시킨다.

6. 문제점 및 대책

6. 1. 저작권 문제점 및 대책

저작권 문제는 원문 DB를 구축할 시 고려해야 할 문제 중 하나이다. 과거의 인쇄문헌과는 달리 전자문헌은 보다 신속하고 쉽게 복제될 수 있으며, 또한 네트워크를 통해서 광범위하게 전송 또는 배포될 수 있으므로 저작물 이용자와 저작권자의 이해조절이 더욱 요구되고 있다.

도서관의 경우에는 저작권법 제 28조(도서관 등에서의 복제 등)에서 규정한 범위내에서 복제할 수 있으며 저작권법 제28조는 아래와 같다.

① 도서관 및 독서진흥법에 의한 도서관과 도서 · 문서 · 기록 그 밖의 자료(이하 "도서관등"이라 한다)를 공중의 이용에 제공하는 시설 중 대통령령이 정하는 시설(이하 "도서관 등"이라 한다)에서는 다음 각호의 1에 해당하는 경우에 보관된 자료를 사용하여 저작물을 복제할 수 있다.(1999. 12. 개정)

1. 조사 · 연구를 목적으로 하는 이용자의 요구에 따라 공표된 도서 등의 일부분의 복제물을 1인 1부에 한하여 제공하는 경우(1999.12. 개정)

2. 도서관 등이 자료의 자체보존을 위하여 필요한 경우

3. 다른 도서관 등의 요구에 따라 절판, 그 밖에 이에 준하는 사유로 구하기 어려운 저작물의 복제물을 보존용으로 제공하는 경우

② 도서관 등은 컴퓨터 등 정보처리능력을 가진 장치를 통하여 당해 시설과 다른 도서관등에서 이용자가 도서 등을 열람할 수 있도록 이를 복제 · 전송할 수 있다. 이 경우 도서관등은 이 법에 의하여 보호되는 권리를 위하여 필요한 조치를 하여야 한다.(1999.12 본항 신설)

이에 본 도서관에서는 '99년도 후기 학위논문 수령분부터 학위논문 제출자에게 저작권 사용동의서를 받고 있으며, 이전의 학위논문에 대한 대책으로는 홈페이지에 저작권 사용 동의서 양식을 올려서 졸업생들에게 동의를 받을 예정이다.

6. 2. 검색상의 문제점 및 대책

로 입력했으나, IR Engine를 예산문제로 구입하지 못한 관계로 초록과 목차의 키워드 검색이 제공되지 못했다. Oracle version에서도 module을 개발해서 키워드를 추출할 수는 있으나 건수가 많아지면 검색속도가 현저히 떨어지므로 IR Engine의 보완이 필수적이었다. 앞으로는 서울시립대학교 종합정보화 2차년도 사업 수행 중 원문정보시스템의 보완으로 RDBMS도 Oracle 8로 Upgrade하고 IR Engine 도 구매하여 초록 · 목차 뿐 아니라 원문의 전문을 대상으로 하는 검색도 제공할 예정이다.

7. 결　　론

전자도서관에서는 메타데이터인 서지정보를 제공하던 문헌정보시스템과 더불어 본문의 내용을 제공하는 원문정보서비스는 이용자에게 좋은 서비스를 제공하기 위하여 무엇보다 중요하다.

본 도서관에서는 서울시립대학교 종합정보화 1차년도 사업 중 원문데이터베이스 구축사업으로 먼저 본교 석 · 박사 학위논문을 원문데이터베이스 구축하였으며, 2차년도에는 본교 부설 발간자료를 대상으로 구축할 예정이다.

본 도서관에서 원문 DB 구축시 사용한 XLX방식은 각종 워드파일을 한꺼번에 자동으로 변환시킬 수 있고, 파일의 크기가 작아 저장과 전송에 유리하고, 전용뷰어가 TIFF파일을 지원하는 장점이 있다. 하지만 PDF방식과 달리 범용성이 낮아 다른기관의 원문 DB를 같은 뷰어로 볼 수 없는 단점이 있다.

SGML은 구조적인 문서의 표현이 가능하고 사용자정의에 따라 다양한 형태로 문서를 작성할 수 있으므로 여러 나라에서 국제표준으로 인정하고 있다. 이의 적용에 대한 연구가 활발히 진행되고 있으며, 따라서 앞으로 SGML이 표준이 될 가능성이 매우 높다. 그러나 현재에는 SGML적용시 고려해야 할 사항이나 개발해야 할 기능이 많은 관계로 완벽한 처리는 기간이 필요하며, SGML로 DB를 구축하는데는 일반적으로 SGML개발도구가 비싸 구축비용이 많이 든다.

본교에서는 XLX포맷으로 원문DB를 구축하였으며, 앞으로는 원문정보시스템의 보완으로, XLX파일뿐 아니라, Web 브라우저 기반으로 이용자 인터페이스가 용이한 PDF파일로도 서비스를 제공할 예정이다.

參 考 文 獻

1. 국회도서관 외, "국가전자도서관 구축 기본계획", 1997.
2. 김길중, "보존문서의 디지털화 방안", 디지틀 도서관 '99 겨울호(1999). pp.37-42.
3. 김성혁, "디지털도서관의 문헌 특성 및 관리", 한국문헌정보학회지 제 31 권 제 1호(1997). pp.53-70
4. 김용철, "가상대학 운영과 디지털 도서관 구축에 관한 연구", 디지틀 도서관 '98 여름호(1998). pp.67-83
5. 김준수, "한국방송대학교 전자도서관 구축사례 : 원문정보시스템을 중 심으로", 디지틀 도서관 '98 여름호(1998), pp. 114-123.
6. 박재영, "전자도서관 모형 및 구축현황에 관한 연구", 연세대학교 대학 원석사학위논문, 1995.
7. 방준필, "대학도서관의 학위논문 전문DB 구축방안", 한국비 블리아 제9집 (1998). pp.39-52.
8. 백성규, "디지털도서관의 원문이미지정보 시스템 구축에 관한 연구", 단국대학교 경영대학원 석사학위논문, 1997.
9. 안현수, "전자출판을 위한 파일포맷", 디지틀 도서관 '97 여름호(1997). pp.96-103.
10. 유사라, "정보화사회와 도서관 정보네트워크", 나남출판, 1996.
11. 이인수, "전자도서관을 위한 논문 원문 검색시스템의 구현", 명지대학 교 산업기술대학원 석사학위논문, 1995.
12. 전자신문, "어도비 국내 전자문서 시장공습", 1999. 11. 19.
13. 정보통신신문, "텍 플러스 〈셈틀로 미디어〉", 1999. 5. 31.
14. 정상조, "도서관에서의 저작권", 국회도서관보 제33권 6호(1996)pp. 10 -13.
15. 한국파일링, "디지탈 도서관 시스템 구축기술", 한국파일링, 1996.
16. 한양시스템, "원문정보 데이터베이스 구축을 위한 JET DOCUMEN T",디지틀 도서관 '98 겨울호(1998). pp.128-130.
17. 홍재현, "네트워크 환경에서의 디지털 복제와 공정사용 법리 적용의 문제점", 한국문헌 정보학회지 제31권 제4호(1997).pp.139-164.

18. 황찬현, "전자도서관과 저작권", 국회도서관보 제33권 6호(1996), pp. 15-29.
19. http : //www.hanyang.co.kr
20. http : //www.nanet.go.kr/nal/3/3-1-4/nal97023.htm