비정형보고서의 변환 모델 및 시스템 구축에 관한 연구

A Study on the Conversion Model and Implementation of the Systems for Unformatted Reports

허태상, 최기석*, 주원균**
한국과학기술정보연구원

Huh Tae-Sang, Choi Ki-Seok*, Joo Won Kyun**
Korea Institute of Science and Technology Information

요약

최근 급속한 기술의 발전과 더불어 연구개발 산출물도 상당량 발생을 하고 있다. 종전의 데이터베이스 구축 방식으로는 방대한 산출물을 디지털 아카이브하기엔 많은 시간과 비용을 초래한다. 또한 기술 및 지식사회가 되어감에 따라 국가R&D보고서 콘텐츠도 다양한 전문서비스가 가능하도록 웹서비스 개발이 요구되고 이를 충족시키기 위해 전문서비스가 가능한 E-Book 형태로 가공하는 절차가 필요하다. 본 논문에서는 국가R&D보고서의 E-Book을 정의하고 효율적인 가공절차를 수용하는 변환 모델 및 관련 시스템 구축에 대해 논의한다.

Abstract

Large amounts of research results are being produced with the rapid growth of science and technology. The current method for constructing the database was way too time and cost consuming. In the currently upcoming technology and knowledge society, web service is required to provide the various professional service for national R&D reports. A process to manufacture E-Book for professional service is also required. This research shows the conversion models with efficient manufacturing process and the systems related to the development through defining the National R&D reports.

Ⅰ. 서론

국가연구개발사업 규모가 커짐에 따라 디지털아키이빙되는 연구보고서양도 급격히 늘고 있다. 종래방식은 수작업으로 연구보고서 서지정보를 입력을 하였으며, 과제정보에 해당하는 정보에 대해서는 관련 기관 이외에는 별도의 검증방법 조차 없어, 데이터 오류에 대한 개연성이 존재하였고 각 부처별로 관리되고 있는 연구보고서 양식은 표준화가 되어 있지 않아서, 일괄적이고, 체계적인 정보생성이 어려웠다. 서지정보 중 일부정보(보고서명, 연구기간, 저자명, 주관연구기관, 발주기관, 발주부처, 보고서번호, 기술분류)는 연구과제정보와 동일하거나 변형의 성격을 가지고 있어 과제정보를 활용하면, 보다 정확하고 충실한 서지정보를 생성시킬 수 있다.[1] 또한 전자문서의 전문서비스를 위해 원문 추출을 통한 서지정보 생성/검증과 목차링크를 자동 생성 및 검증할 수 있어야 하고 다른 파일 포맷과는 상이하게 국내 소프트웨어 포맷인 hwp는 버전에 따라서 문서 변환 시 발생할 수 있는 폰트, 이미지, 레이아웃의 변형 없이 원본상태로 유지할 수 있어야 한다. 또한, 기존의 C/S 환경에서 원문에 대한 수정, 변형 등이 일어날 경우 책임이 불명확해지는 문제점이 있어 이를 해결하기 위해서는 문서의 맥락정보와 문서내용이 표준화된 생산, 유통과정을 준수해야 한다.[2][3] 본 연구는 기존의

비정형 보고서 변환을 관련 연구를 통한 개선 모델과 그 시스템 구축을 중점으로 논의한다.

Ⅱ. 관련연구

1. E-Book 콘텐트 제작

E-Book의 광의적 의미로는 모든 형태의 디지털 매체를 이용해 지식과 정보를 출판하는 것으로, CD-ROM, 인터넷, PC통신을 이용한 온라인 출 등을 포괄하는 개념이고, 협의적 의미로는 도서로 간행 전, 후의 저작물 내용이 디지털 데이터로 전자적 기록매체, 저장장치에 수록되고 On/Off-line으로 그 내용을 인지할 수 있는 것을 말한다. 현재 사용되는 표현으로 E-Book, 전자북(electronic Book), 웹북(Web Book), 전자카탈로그(electronic catalog), 웹카탈로그(Web catalog), 파일북(File Book), 온라인북(On-line Book)는 광의적으로 같은 의미를 같기도 한다.[4][5] E-Book 콘텐츠의 제작은 크게 세가지 솔루션을 바탕으로 이루어진다. 전문서적, 비전문서적 또는 애니메이션 등을 E-Book으로 제작하는 대상이라 할 수 있고 불법복제 방지기술을 포함할 수 있어야 한다. 보통 E-Book을 볼 때는 전문 뷰어(Viewer)가 일반적이나 요즘은 전문뷰어 없이 실행파일 형태로 제작이 가능하다.[6] 비전문서적은 단순 텍스트로 되어 있어 XML(eXtensible Markup Language)을 이용하여 제작하는 것이 일반적이다.[10] 문서의 논리구조를 정의하고 재사용성이 종전의 HTML방식 보다는 뛰어난 언어이기 때문에 미국의 OeBF(Open eBook Publication Structure, 일본의 JEPA, 그리고 우리나라의 EBK(e-Book of Korea) Consortium는 XML기반의 표준안을 채택하여 사용하고 있으며, 국내 몇 몇 제품은 XML을 지원하는 솔루션을 출시하는 경우도 있다.[7][8][9]

전문서적은 일반적으로 이미지를 많이 포함하고 있으며, 편집형태를 책자와 동일하게 유지해야 함으로, 이경우에는 PDF(Portable Document Format) 방식을 채택한다. 레이아웃을 유지하고, 비트맵 그래픽과 벡터 그래픽 지원, 멀티미디어적 표현 가능한 포맷이라 할 수 있다. 애니매이션은 적은 용량의 역동적인 속성으로 저작도구의 하나인 플래쉬(Flash)로 표현하는 것이 가능하고, 문자적 의미전달 보다는 멀티미디어적 요소의 의미전달이 강력한 장점이 될 수 있다. 최근에는 플래시페이퍼 기술이 개발되어 PDF문서처럼 서체와 레이아웃, 이미지 등을 처리하기도 한다.

국가R&D보고서 콘텐츠는 제작 관점에서는 전문서적으로 분류될 수 있고, 대용량의 학술정보를 효과적으로 전달할 수 있다. 또한, 플래쉬 기반의 F-Book은 HTML의 기본을 유지하면서, 하이라이트 표시, good design 스킨, 리사이징(resizing) 등의 장점을 가지고 있어 차선의 콘텐츠 제작도구라고 할 수 있다.[10]

2. E-Book 솔루션 업체

해외로는 PDF 형식을 지원하는 Adobe사의 Acrobat Reader, XML 형식을 지원하는 Microsoft사의 MS-Reader 등이 대표적이다.[11] 국내에서는 다양한 콘텐츠를 지원하는 E-Book 솔루션들은 종류도 매우 다양하고 기술 수준 또한 세계수준이라 할만하다. 국내 주요 E-Book 솔루션 업체로는 이앤아이월드, 아이비엘, 에프데스크, 유니닥스, 디지토미, 엔씨투, 휴먼드림, 레몬북, 비엔씨, 이피루스, 소프트북, 아이하우스, 디코시스, I&T System, 컨텐츠 밸리, VRPhoto, 아이카탈로그 등이 있다.

3. 국가R&D보고서 PDF 제작

PDF 자체의 여러 가지 속성이 문서의 교환을 위한 전자문서 포맷으로서 장점을 가지고 있다. PDF로 변환된 파일은 사용자의 PDF 뷰어을 통해 어떤 OS나 소프트웨어의 설치없이 원본을 열람할 수 있다. 주요 속성으로는 레이아웃, 스타일, 서체정보 등을 포함하고 있으며, 문서열람, 내용 추출, 인쇄 출력에 대한 제어정보를 포함하고 있어 보안적 속성도 내포하고 있다[12].

시스템 구축을 위해 PDF 변환기를 선별은 시스템의 안정성, PDF 포맷의 일반성 및 확장성, 유지와 보관의 용이성을 고려해야 한다. 이를 위해 가능 벤더들을 분석하자면, 먼저 PDF의 기술의 창시적 위치에 있는 Adobe PDF와 그 외 기술 공유를 토대로 PDF를 만드는 유사 PDF 제품군간의 기술 비교가 필요하고 내용은 아래[표 1]과 같다.

표1. Adobe PDF 제품군과 유사 PDF 제품군 비교

항목	Adobe PDF 제품군	유사 PDF 제품군
PDF 포맷	- PDF 1.3 ~ 1.6까지 - ISO 인증 장기보관 문서 포맷인 PDF/a 지원 - PDF/X-1a, PDF/X-3	- PDF/A-1
변환 대상	- Office제품군 - 아래한글 - 훈민정음 - jpg, bmp, gif, tiff, png - html, WEB-URL, eml	- Office제품군 - 아래한글 - 훈민정음 - jpg, bmp, gif, tiff, png
Plug-in 기능	- PDF병합, 워터마크, 북마크, 보안설정 - 파일삽입, 텍스트 추출 - XMP 메타데이터 삽입	- PDF병합, 워터마크, 북마크, 보안설정 - 파일삽입, 텍스트 추출 - 메타데이터 삽입
자동 변환	감시폴더 자동변환(서브폴더 포함)	- 감시폴더 자동변환(서브폴더 포함)
로그 관리	실시간 변환진행 상태 정보 제공	- 변환 결과 로그 제공
PDF 생성	- 최적화된 자동 PDF생성 서비스	- 최적화된 자동 PDF생성 서비스
배포 옵션	- 유동적인 배포 옵션 제공 - 네트워크감시폴더 - 전자메일 - 웹 인터페이스 - Java Bean API	- 유동적인 배포 옵션 제공 - JAVA, VB API 제공 - 웹 인터페이스
적용 범위	- DB 어플리케이션과 연동하여 생성하는 1회성 문서	- 문서공유, 장기보관, 불특정 다수 배포 서비스 등
문서 보안	- 기본적인 패스워드 보안 셋팅 및 APS를 이용한 보안 기능 확장 - 문서내의 텍스트 및 그림의 Capture 방지 설정 - Adobe Reader 에서 문서의 인쇄 및 저장 기능 비활성화 설정	- PDF문서 암호화, 인쇄, 내용복사 제어 기능
표준 기반	- J2EE기반 어플리케이션, HTTP(S), SMTP, 다양한 DB지원	- J2EE기반 어플리케이션, HTTP(S), SMTP, 다양한 DB지원
FAST Web View	- PDF버전에 관계없이 빠른 웹보기 기능	- PDF/A에 대한 빠른 웹보기 기능
장점	- PDF 창시 기업임에 따른 인지도, 국제적인 표준 제시, 안정적 운영기반	- 요구사항에 대한 커스터마이징 제공

또한 Adobe PDF 제품군은 문서특성 컨트롤 부문에서 파일사이즈, 파라미터가 최적화되어 있고, 기존정보 보존도 링크, 태그, 북마크 등과 같은 특정파일 포맷의 기능을 보존하여 변환이 가능하다.[1)]

표 2. 국가R&D보고서 PDF 보안 및 속성

보안	구분	속성
설명	PDF변환 소프트웨어	Acrobat Distiller 8.0.0(windows)
	PDF버전	1.6(Acrobat 7.x)
	빠른 웹 보기	예
문서 보안	보안방법	암호보안
	열수 있는 버전	Acrobat5.0이상
	문서열기암호	암호보안
	권한암호	예
	문서변경	허용안됨
	시작필드 채우기 또는 서명	허용안됨
	페이지 추출	허용안됨
	암호화 레벨	128비트RC4
문서 제한 요약	인쇄	허용
	문서 어셈블리	허용안됨
	엑세서빌러티를 위해 내용복사	허용안됨
	주석달기	허용안됨
	양식 필드 채우기	허용안됨
	서명	허용안됨
	템플릿 페이지 작성	허용안됨

앞서 관련 연구를 토대로 한국과학기술정보연구원(KISTI ; Korea Institute of Science and Technology Information)에서는 2005년부터 연구보고서를 PDF 포맷으로 대국민 서비스를 하고 있었으며, 2008년에는 국제표준화기구(ISO, International Organization for Standardization)에서 대표적인 전자문서 파일 표준 포맷인 PDF를 국제표준(ISO 32000-1)로 지정하였다.[2)]

1) http://www.adobe.com

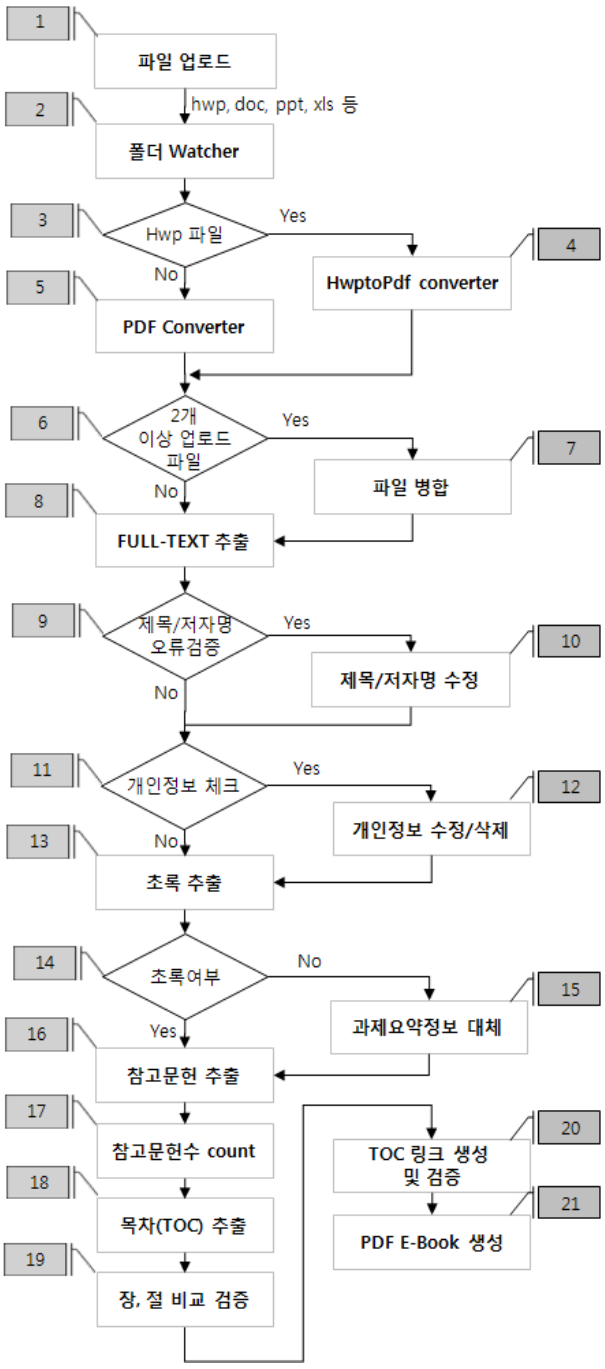
이에 KISTI에서는 2008년부터는 PDF 보안 및 속성을 [표 2]로 정의하여 콘텐츠 정보를 변경/적용하였다.

Ⅲ. 비정형보고서 변환방법 및 구축

본 연구는 국가연구개발사업의 성과물인 연구보고서는 다양한 문서파일 포맷으로 생성된다. 비정형 보고서 변환방법은 변환된 전자문서 내 원문내용을 추출한 후 해당 서지정보를 자동생성함은 물론, 등록 정보의 오류를 체크하고, 개인정보를 원문 생성 룰에 의해 자동 삭제/수정하며, 추출된 목차정보와 목차링크를 생성, 검증하는 절차를 포함한다. 본 프로세스는 연구보고서에 대해서 데이터베이스를 구축 시 서지정보를 보다 정확하고 쉽게 등록받을 수 있고, 전자 문서내의 목차 정보를 추출하여 비교 알고리즘을 통해 다량의 연구보고서 전자원문을 전문서비스가 가능한 TOC를 생성할 수 있다. 과거 연구보고서 정보 생성 시, 비전문인력이 연구보고서 정보를 수동 입력 인터페이스를 이용하여 입력하고, 수작업으로 목차정보와 전자문서 내 링크를 만들어서 연구보고서 정보 오류의 가연성이 있었다.

본 연구에서는 과거 발생했던 문제점을 보완하였고, 외부 기반정보와의 연계와 자동 영문변환 기능을 통해 사용자 입력 오류를 최소화하고, 전자문서(HWP, DOC, PPT 등)로 작성된 비정형 보고서를 자동 변환기 및 비교 알고리즘을 통해 TOC 링크정보를 가지는 전자문서(PDF)를 효과적으로 구축할 수 있는 변환방법 및 구축에 대해서 논의한다. [그림 1]은 비정형 보고서(hwp, doc, ppt, xls 등)의 소스 전자문서 가공처리 방법에 대한 그림이다. 본 처리방법은 보고서 DB구축에 있어 가장 시간과 비용이 많이 드는 절차에 대한 자동화 기법을 도입하여, 효율적으로 전자문서를 가공하는 데 그 목적을 가지고 있다. 또한, NTIS와 연계를 통한 일부 과제정보 기반의 서지정보를 가져옴으로써 과거 수작업 입력 시 문제점으로 제시되었던 누락정보를 보완할 수 있다.

2) 전자신문 컴퓨팅, 2008.07.07 "PDF, ISO 국제표준으로 지정"

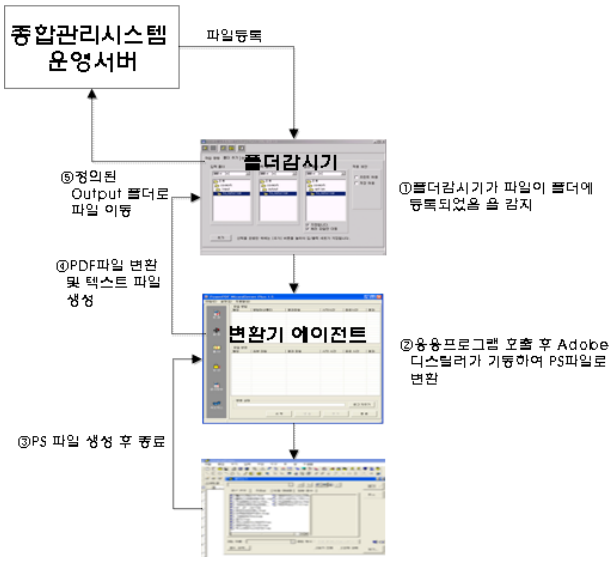


▶▶ 그림 1. 비정형 보고서 원문 가공처리 절차

본 원문 가공처리 절차는 보다 정확한 원문상태를 유지하고, 원문 내 개인정보의 유출을 방지할 수 있다. 또한 장문의 인쇄물을 전자문서화할 때 사용자 편의성인 본문 내 검색과 목차링크를 제공하여 전문의 빠른 보기가 가능해진다. 검색 서비스에서도 이를 활용하여, Full Text 원문 검색과 초록, 참고문헌에 대한 전문 사용자 검색조건을 만족시켰으며, 과제정보를 기반으로 하기 때문에 타 성과물 시스템과의 연계가 가능하다.

표 3. 비정형보고서 원문 가공처리 절차 설명

순번	내용	설명
1	파일업로드	보고서 파일 업로드
2	폴더 Watcher	폴더 Watcher에서 업로드 확인
3	HWP파일 체크	소스 파일이 HWP인지 확인해서 'Yes'면 HwptoPdf Converter에서 처리
4	HWP to PDF Converter 처리	한글문서 처리
5	PDF Converter	PDF로 변환
6	파일 개수 체크	변환 파일 개수를 체크해서 파일개수가 2개이상이면 파일병합 처리
7	파일병합	2개 이상의 PDF파일을 합치기
8	Full-Text 추출	원문 내 Text 추출
9	제목/저자명 오류검증	제목/저자명 오류 체크해서 오류 발생 시 수정처리
10	제목/저자명 수정	제목/저자명 수정
11	개인정보 체크	개인정보(주민번호, 계좌번호, 카드번호) 유무를 체크해서 'Yes'면 수정/삭제
12	개인정보 수정/삭제	개인정보 수정/삭제
13	초록추출	Full-Text에서 초록 추출
14	초록여부체크	초록 여부를 체크해서 'No'면 과제요약정보로 대체
15	과제요약정보 대체	과제요약정보 대체
16	참고문헌 추출	참고문헌을 Full-Text에서 추출
17	참고문헌수 Count	추출된 참고문헌을 Count
18	목차(TOC)추출	Full-Text에서 TOC를 추출
19	장,절 비교 검증	원문과 장절을 비교
20	TOC 링크 생성/검증	TOC 링크를 생성 후 원문내에서 링크를 직접 확인
21	PDF E-Book 생성	PDF E-Book을 생성



▶▶ 그림 4. PDF 변환모듈

[그림 4]는 PDF 변환 모듈로써, [표 3]의 비정형보고서 처리절차 1에서 5까지의 프로세스를 Window OS에서 동작하도록 개발한 그림으로, 파일등록 시 Watcher (폴더감시기)를 통해 PS 파일로 변환 후 PDF 변환을 과정을 통해 종합관리시스템 서버로 텍스트 파일을 생성하여 종합관리시스템으로 텍스트 파일을 제공하도록 구현하였다.

Ⅳ. 결론

본 연구에서는 비정형화된 연구보고서 전자문서를 PDF 변환 과정 후, 원문 내 정보 추출과 검증을 통해 서지정보를 자동 생성, 검증하고, 전자문서의 유통 상 개인정보 보호를 위해 주민번호,계좌번호, 카드번호를 검증하는 기능을 제공한다. 또한 많은 페이지량을 가지는 연구보고서에 대해 사용자의 용이한 접근성을 제공하는 목차 정보링크 생성 시 많은 부분 자동화 기법을 도입함으로써 방대한 양의 전자문서를 빠른 시간에 정확하고 효율적으로 구축할 수 있는 기반 모델을 제시하였다.

본 연구를 통해 보고서 서지정보의 오류의 개연성을 많은 부분 제거를 하였으며, 원문 내 개인정보의 유출 우려도 사라질 것으로 기대된다. 종 래 시스템으로 전자원문을 한 건 처리하는 데 드는 시간을 기준으로 비교해 볼 때 본 시스템은 가공처리하는데 30% 정도의 시간적 절감

효과를 가진다고 볼 수 있고, 직접적으로 비용적인 면에서도 많은 절감효과를 가져왔다고 볼 수 있으며, 이는 성과관리전담기관으로써 연간 3만 여건 이상 등록되는 보고서의 고품질의 가공처리가 가능하리라 판단된다.

■ 참 고 문 헌 ■

[1] 허태상, 최기석 "국가연구개발보고서 관리시스템의 개선방안에 대한 연구", 한국콘텐츠학회 06 추계학술대회논문집, 10, pp.693-697, 2006, Nov.

[2] Joseph Marin and Julie Shaffer, "The PDF Print Production Guide", Graphic Arts Technical Foundation, 2003.

[3] 남영준, 정의섭, 유재영, 조현양, "웹기반의 전자원문 관리 시스템에 관한 연구", 한국비블리아학회지, vol.16, no.2, pp.139-156, 2005

[4] 김경옥, 김성혁, 임순범, 최윤철, "eBook 메타데이터 비교 및 한국전자책표준의 메타데이터 개발 ", 한국전자거래학회, 01 International Conference CALS/EC KOREA, pp.511-521, 2001. Aug.

[5] 손원성, 고승규, 이경호, 김재경, 김성혁, 임순범, 최윤철, "한국전자책문서표준(EBKS)의 개발", 정보관리학회지, vol.18, no.2, pp.255-272, 2001

[6] 김경일, "전자책 콘텐츠 산업의 현황과 전망에 관한 연구", 디지털콘텐츠학회 논문지, vol.7, no.2, pp.269-284, 2006

[7] EBK(e-Book of Korea)Consortium, "A Study of Korea Standardization of eBook documents", Technical Report, 2001

[8] Open eBook Publication Structure 1.0. Open eBook Forum(OEBF), http://www.openebook.org

[9] Japanes Electronic Publishing Association(JEPA), http://www.jepa.or.jp.

[10] 나재무, 백혜선, 이광재, 이정훈, 하동훈, 이은정, "전자도서관을 위한 전자책 교환 시스템 개발", 한국정보과학회, 2002봄 학술발표논문집(A):Proceedings of The 29th KISS Spring Conference, pp.808-810, 2002. Apr.

[11] 한국소프트웨어진흥원, "디지털컨텐츠 중장기 육성전략 연구보고서", 2002. Dec.

[12] 이재득, "웹을 기반으로 한 PDF 출판 솔루션에 관한 연구", 산업경영시스템학회지, Vol.47, no.3, pp.184-189, 2005.