웹 환경의 EDMS를 위한 전문 정보시스템 구축

Construction of Web-Based EDMS Supporting Full-Text Information System

김성도(S.D. Kim) 기술정보서비스팀 선임기술원
김극희(G.H. Kim) 기술정보서비스팀 기술원
이은철(Y.C. Lee) 기술정보서비스팀 책임기술원, 팀장

본 연구는 최근 들어 각광을 받고 있는 전자문서관리시스템(EDMS)의 주요 기능 중에서 문서의 공유와 문서 형식(file format)의 변환을 통한 전자문서 관리의 효율성 향상을 위해 어떻게 전문 정보를 구축할 것인가에 초점을 맞추고 문서의 통합 기능을 제공하는 웹 기반의 EDMS를 설계 및 구현하였다. 이를 위하여 인터넷상에서 현재 표준 문서형식으로 사용 중이거나 개발중인 전자문서 포맷들의 특징을 비교, 분석하고, 정보 시스템의 관점에서 전문정보를 어떤 문서 포맷으로 처리하는 것이 적합한가를 검토하였다. 그리고 기존의 문헌을 전자화 하고 새로운 전자문헌을 제작하기 위한 보다 효율적인 방안 제시를 위하여 전문(全文) 정보를 PDF 문서로 제공하는 정보 시스템을 구현하였다.

I. 서론

근래에 외서 인터넷의 등장, 하드웨어/그룹웨어 기술의 발달 그리고 이미진이나 COLD(Computer Output to Laser Disk) 등의 시스템 사용이 확산됨에 따라 조직내에서 생산되는 정보의 대상은 이전의 종이 문서에서 전자 문서의 형태로 급격하게 바뀌고 있으며, 이는 곧 전자문서관리가 중요한 문제로 대두됨을 뜻한다. 전자문서관리시스템(Electronic Document Management System: EDMS)은 문서 자체가 종이에 기록된 형태가 아닌 컴퓨터 기반에서 생성된 모든 종류의 전자적 문서뿐 아니라 E-mail과 팩스 등 외부에서 유입되는 문서를 통합 관리해 주는 시스템이다.

EDMS의 가장 큰 특징은 문서를 종이가 아닌 파일로 전자화 해서 검색하거나 주고받을 수 있다는 점이다. 근래에는 전자결제와 워크플로 시스템의 도입으로 문서의 종류가 종이형태보다는 거의 워드프로세서 파일로 존재하게 되었는데, 특히 과학기술분야에서 논문이나 기술문서를 전문(full-text)으로 제공하는 경우 조직내의 문서 공유와 문서 표준화를 지향하기 위하여 다양한 문서 형식을 변환하여 원문 정보를 구축하고 있다. 따라서 원문을 처리하는 EDMS에서는 문서의 효율성 향상을 위해 정보를 신속하게 변환하고 저장받음으로 데이터베이스를 구축할 수 있는 방법을 찾는 것이 중요하다.
한 문제가 되고 있다.

본 연구는 웹 기반의 EDMS를 위한 전문 정보의 구축을 위해서 현재 인터넷상에서 표준 문서형식으로 사용 중이거나 개발중인 전자문서 포맷들의 특징을 비교, 분석하고, 정보 시스템의 관점에서 전문정보를 어떤 문서 포맷으로 처리하는 것이 적합한가를 검토하였다. 그리고 기존의 문헌을 혼합하고 새로운 전자문헌을 제작하기 위한 보다 효율적인 방안 제시를 위하여 한국전자정보연구원의 연구문서 집합관리 방안으로서 연구원내의 그룹웨어와 연계된 전자문서관리 시스템을 구현하였다.

본 논문의 구성은 다음과 같다. 제 II장에서는 전문 정보 촉적을 위한 전자문서 포맷의 발전과정과 현황 및 장단점을 비교, 분석하였고, 제 III장에서는 기존 정보관리시스템에서 이용되고 있는 전자문서 포맷과 웹 기반 정보관리시스템에서의 전자문서 포맷을 살펴보고 이들 전자문서 포맷을 이용하여 EDMS 시스템에서 활용 가능한 문서 포맷으로의 변환을 실험하였다. 그리고 제 IV장에서는 실제 대상 문헌을 PDF(Portable Document Format) 전문정보로 처리하는 Web 기반의 EDMS를 설계 및 구현하였다.

II. 전자문서 포맷을 이용한 전문 정보의 촉적


그러나 아날로그 정보의 디지털 정보로의 변환, 촉적과 관리비용은 전통적인 방식의 인적·물적 시스템을 구축하는데 드는 비용보다 훨씬 높아지게 되었다. 특히 데이터베이스 구축을 위한 문서 전체의 데이터 인코딩에는 많은 비용과 시간이 소요되며 따라 전문 정보를 보다 효율적이고 저비용으로 촉적할 수 있는 인코딩 방법의 모색이 필요하게 되었다. 이러한 인코딩 방법으로 텍스트 기반의 전문 정보 촉적하기 위한 ASCII 기반 인코딩, 스키어를 이용하여 전문을 이미지화 하는 이미지 기반 인코딩, 그리고 웹의 등장으로 주목받기 시작한 Markup 기반 인코딩 등이 주를 이루게 되었다[3, 12]. 그러나 전문 정보가 초기 생성단계부터 디지털 정보로 이루어지지 않는 이상 이러한 인코딩 방법들은 의에 섬도 많은 시간과 비용이 소요되므로 가장 효율적인 인코딩 방식이 요구된다.

최근 들어 많이 이용되고 있는 인코딩 방식으로 PDF 전자문서 포맷을 들 수 있다. 일반적으로 워드프로세서로 작성된 정보는 단순한 텍스트뿐만 아니라 도표, 그림이 포함된 복합 문서 형태로 이루어져 있다. 기존의 텍스트 내지 클라이언트/서버 기반의 정보 시스템에서는 워드프로세서로 작성된 원문을 그대로 저장할 수 있으나 웹을 기반으로 하는 정보 시스템에서의 문서 포맷은 웹 환경에 적합한 표준화된 전자문서 포맷을 유지하는 것이 중요하다. 이미 웹에는 HTML(Hyper Text Markup Language)이라는 표준화된 전자문서 포맷이 존재하고 있다. 하지만 HTML은 매우 단순하여 복합문
서를 처리하는 것이 어려울 뿐만 아니라 박대한 수 작업과 시간이 소요된다. 더욱이 복합문서를 이미지 기반으로 인코딩하는 경우에는 많은 데이터 저장 공간이 요구될 뿐만 아니라 정보의 재활용도 없어서 번거로워지게 된다. 이러한 두 가지 전자문서 포맷의 단점을 최대한 줄이기 위해 PDF이다.

한편, Markup 기반 인코딩 방법으로 기존 HTML의 단점을 보완하기 위한 새로운 문서 포맷에 대한 연구가 활발하게 이루어지고 있다. SGML (Standard Generalized Markup Language)과 XML (eXtensible Markup Language)이 그 대표적인 주로 저작 및 변환 툴, 그리고 브라우저에 대한 연구가 진행 중이어서 이를 활용한 전문정보의 구축에 증가될 것으로 예상된다[7].

전문 정보 구축을 위한 문서 포맷의 선택은 해당 시스템이 어떠한 형태의 데이터를 다루는 정보 시스템이냐에 따라 어느 정도 좌우된다. 대체적으로 적합한 문서 포맷의 선택이 정보 구축 및 관리 비용의 절감과 더불어 정보 공유와 전달을 통한 정보 활용도를 높이는 가장 핵심적인 요인이 되므로, 다음 III장에서는 정보 시스템에 필요한 각 전자문서 포맷의 현황 및 장단점과 변환 방법을 비교 분석해보고자 한다.

Ⅲ. 정보관리시스템에서의 전자문서 포맷

호스트 또는 클라이언트/서버 기반의 정보관리 시스템에서는 전문 정보를 축적하기 위한 인코딩 방법으로 주로 ASCII 기반이나 이미지 기반의 전자문서 포맷이 가장 널리 이용되고 있다.

이에 반해 웹을 기반으로 하는 EDMS 형태의 정보관리시스템에서는 HTML과 같은 Markup 기반의 인코딩 방법이 주로 이용되고 있다. 따라서 기존의 정보 시스템을 웹 기반으로 전환하고자 하는 경우 Markup 기반으로 전문 정보를 구축하는 것이 보편적인 입장일 수도 있다. 하지만 기존 ASCII 기반이나 이미지 기반의 전문 정보를 Markup 기반으로 변환하는 것은 많은 인력과 비용이 소요되는 방대한 작업이다. 그 예로 최근 미국 국가도서관에서 MARC을 이용하여 총정의 원문 정보를 SGML 문서 포맷으로 변환하기 위한 방법[12] 등이 연구되고 있으며 변환에 요구되는 시간과 비용이 상당할 것으로 추정되고 있다.

따라서 기존의 정보 시스템을 웹 기반의 정보 시스템으로 전환하고자 하는 경우 Markup 기반이나 아닌 새로운 전자문서 포맷이 요구되고 있으며, 이러한 문서 포맷으로 PDF가 주목받고 있다. 특히, PDF는 기존 ASCII나 이미지 문서 포맷뿐만 아니라 워드프로세서로 작성된 원문정보의 처리도 매우 용이하여 웹 기반의 정보 시스템에서 새로운 문 서 포맷으로 유용하게 이용할 수 있게 되었다. 다음에서는 기존 정보관리 시스템과 웹 기반 정보관리 시스템에서의 각 전자문서 포맷별 문서 구축에 대한 장·단점을 비교 분석하기로 한다.

1. 기존 정보관리시스템에서의 전자문서 포맷

가. ASCII 기반 문서 포맷

ASCII 기반 시스템은 기존의 모든 텍스트를 키보드로 입력하여 ASCII 코드로 저장하는 텍스
트 응용 기법을 사용하는 방법과 OCR(Optical Character Recognition) 소프트웨어를 사용하여 ASCII 코드로 변환시키는 방법을 적용한다[10]. ASCII 기반 시스템에서는 문서에서 사용된 용어를 기준으로 하여 전문 검색과 주제 검색이 가능하며, 기존의 색인보다 효율적인 색인을 자동 생성할 수 있으므로 정보 검색의 속도가 향상될 수 있고, 시스템 저장 공간이 절약된다. ASCII 코드의 정보는 텍스트 전체, 또는 각 부분들에 대한 제목과 제목, 출판을 위한 텍스트 편집에 의한 인쇄본과 CD-ROM출판 및 온라인 서비스 등 다양한 용이적으로 활용이 가능하다. 하지만 원문이 ASCII 형태로 구축되어 있지 않을 경우 ASCII로의 변환이 어려우나 불완전한 점이 있으며, 텍스트 구조를 나타내지 못하기 때문에 텍스트에 대한 구조 탐색이 불가능하다는 단점을 갖는다[4].

나. 이미지 기반 문서 포맷

이미지는 1960년대 후반 마이크로 필름을 사용한 사무자동화에서 출발하여 80년대 광 디스크의 발전으로 이용이 가속화 되었다[10, 21], 즉, 종이 문서의 보존, 관리 및 재활용을 위한 광 파울링 시스템에서 주로 이용되어 왔다. 그 예로 LG 상남도 서관의 웹 사이트[25]에서는 잡지류 1,000종과 회의자료 150-200권 등 약 20만 건의 이미지 데이터를 기반으로 서비스를 제공하고 있다[12].

이미지는 저장 메체의 안정성, 관리의 편리성으로 인하여 전문 정보의 완전성(integrity) 보장, 물리적 저장 공간의 절약이라는 장점을 가지고 있다. 하지만 전문 검색과 구조에 기반한 검색이 불가능하며, 시스템에서 대규모의 기록용량이 필요하고 이미지의 삽입, 삭제 등의 조작이 어려워 응용력이 낮다는 단점을 지니게 된다. 또한 이미지의 질적 인(quality) 부분에도 문제가 있다. 따라서 정보의 검색 및 공유, 전달이 중요하게 되는 정보 시스템에서 높은 비용의 균형적이다[3].

2. 웹 기반 정보관리시스템에서의 전자문서 포맷

가. Markup 기반 문서 포맷

웹에서 사용되는 Markup 기반 문서 포맷의 대표적인 형태는 HTML로 HTML은 웹을 위해 개발된 언어이다[8, 14]. 웹이 인터넷의 광도를 바탕으로 대중화된 발전을 이룬것이며, 이의 발전에 가장 큰 공헌을 한 것이 바로 HTML이다. 이것은 HTML이 웹에서 하이퍼텍스트 및 하이퍼미디어 기능을 지원하며, 누구나 사용할 수 있을 만큼 간단하고, 특별한 데이터 탐색을 사용하지 않는 단순한 텍스트이기 때문에 이르게 늘고 사용이 편리하다. HTML은 하이퍼 링크를 통해 다른 문서와 연결이 가능하도록 문서의 논리적 구조를 규정하기 위하여 설계된 일종의 SGML의 문서 형식 정의(Document Type Definition: DPT)이다. 웹 브라우저는 HTML로 작성된 문서를 해독하여 보여주는 역할을 하며 웹 브라우저의 발전에 따라 HTML 기능도 점차 확장되면서 최근에 발표된 사양의 경우 특정 브라우저에서만 지원하는 사례가 늘고 있다. 더욱이 HTML의 표준화는 단말간의 특수적으로 이루어질 전망이다[15-17].

개인 홈페이지 또는 단순한 홈페이지 작업을 위hetic HTML이 매우 편리하여 대용량의 온라인 콘
관에 HTML을 이용하기란 여간 불편하지 않다. 이는 HTML의 기능적인 측면에서 살펴보았을 때 화면상에 보여지는 기능 외에는 별다른 기능을 제공하고 있지 않을 뿐만 아니라 고정된 태그만을 사용하고, 또 페이저 레이아웃 형태를 임의로 지정할 수 없기 때문이다. 이와 같은 단점 중에 많은 부분이 CSS(Cascading Style Sheet)로 해결될 수 있지만, 근본적으로 고정적인 태그셋(tag set)에서 발생하는 문제나 HTML 문서의 구조적 정보를 담을 수 없다는 것은 여전히 개선할 수 없는 문제이다[11, 21].


SGML은 시스템 또는 플랫폼에 따라서 독립적이며 CALS 또는 EC 등과 같은 업계 표준으로 정착하여 많이 사용되고 있다. 하지만 SGML 자체가 위약 복잡하기 때문에 SGML 전체를 지원하는 소프트웨어의 개발이 용이하지 않다. 또한 SGML은 소프트웨어 산업 전반에 걸쳐 널리 사용될 수 있도록 범용 목적으로 만들어져 있어 화학식의 표현이나 기타 다른 특수 용도의 목적으로 사용하려면 그 목적에 맞는 브라우저를 일일이 개발해야 하는 단점을 갖고 있다.

웨이 지금까지 설명하는 데에는 누구나 쉽게 만들고 사용할 수 있는 HTML의 단순함이 그 효율을 하였지만 사용자의 요구가 다양해지면서 이러한 요구를 수용하려는 시도가 바로 XML이 등장하게 된 이유를 하나이다. XML이 등장하게 된 또 다른 이유는 SGML이 너무나 복잡하고 어렵다는 것이다. 즉 HTML과 SGML의 장점을 수용하면서 단점을 극복한 것이 XML이다[11].

XML 스펙을 SGML의 입장에서 살펴보면, SGML에서 거의 사용하지 않는 것은 모두 없이고 끝 필요한 기능만을 수용하였지만 SGML의 중요 한 많은 기능이 그대로 남아 있기 때문에 XML은 SGML의 서브셋(subset)이라고 말할 수 있다. 따라서 SGML을 XML로 손쉽게 변환할 수 있고, 또한 XML을 수정하지 않고도 모든 SGML의 응용에서 사용할 수 있다는 특징이 있다.

나. PDF 기반 문서 포맷

PDF는 Adobe Systems사가 디지털 서류를 전 송하기 위해 개발한 특별한 파일 형식의 하나이 다[9]. 문서를 비롯하여 도표, 그림 등 문서에 포함된 내용을 전용 뷰어(Viewer)인 Acrobat Reader를 사용하여 표시, 출력하는 전자문서 배포용 데이터 형식이다. 또한 PDF는 국제표준기구(ISO)에서 공식 채택되어 전자출판의 표준으로 널리 사용하고 있는 포스트ск립트(postscript) 레이어 기술 언어를 이용한 데이터 파일이다. PDF 문서는 정보 전달의 새로운 파일 형식으로 선진국의 산업계는 물론 공공행정 기관에서도 문서 관리 시스템
의 표준 파일 양식으로 정착되어 사용하고 있으며, 최근 원문을 제공하고 있는 집지출판사의 웹 정보 시스템에서 원문의 전자문서 포맷으로 많이 사용하고 있다.

PDF는 CD-ROM, 인터넷, 인터넷, 전자매일 등의 모든 매체에서 텍스트는 물론 컬러 사진, 음성, 동영상까지도 삽입하여 사용할 수 있고, 인쇄된 기록 자료도 자동 문자 인식으로 글꼴, 단락, 편집 등 원본의 모양을 유지하면서 전문 검색이 가능한 파일 양식이다. 또한 PDF는 시스템 또는 플랫폼과 에뮬레이션에 대해서 독립적이며, 원문 문서의 페이저 포맷 기능과 하이퍼링크 기능, 텔레미디어 데이터 및 전문 데이터 검색 기능, 기밀 문서 보안 장치, 자동 문서 인식기능과 컬랙트 파일 사이즈 등의 특징을 갖는다.

3. 전문정보 구축을 위한 표준 문서 포맷으로의 변환

정보는 처음 생산될 때 특정 EDMS에 적합하게 작성되지 않는다. 즉, 생산되는 정보는 종이 문서, 웹프로세서 문서, 웹 문서 등 다양한 형태로 발생하게 되므로, EDMS에서는 다양하게 입수되는 문서 포맷을 변환하여 이용자에게는 동일한 문서로 제공하는 것이 필요하다. 또한 가급적 표준 문서 포맷을 이용하는 것이 정보 공유와 전달 측면에서 중요하다.

현재까지 전자문서 포맷 관련 기술의 발전으로 보았을 때 XML이 향후 EDMS에서의 표준 문서 포맷으로 활용될 경우 매우 효과적으로 전문 정보를 구축할 수 있을 것으로 기대되지만 그 이전까지 EDMS 시스템 관리자 입장에서는 나름대로 효과적인 문서 포맷을 선택하여 활용하여야 하는 어려움이 있다.

다음은 현재 EDMS의 정보 소스가 되는 웹 프로세서 문서와 종이 문서를 시스템에 적합한 전자문서 포맷으로 변환하는 문서 변환에 대해 살펴보고자 한다.

가. HTML, XML 문서로의 변환

최근 문서 작성 도구에는 아래한글, MS-위드 등 상용 웹 프로세서가 주로 사용되고 있다. 따라서 웹 기반의 EDMS를 위해서는 HTML 에디터나 HTML DTD 스펙에 따라 텍스트 에뮬레이션을 이용하여 HTML 문서로 작성하기 보다는 웹 프로세서 문서를 HTML 문서로 변환하는 방법이 훨씬 편리하다. 변경 방법은 웹프로세서에 포함된 HTML 문서 변환 기능을 이용할 수 있다. MS-위드 97 또는 아래한글과 같은 웹 프로세서에는 웹 문서를 HTML 문서로 자동 변환해 주는 기능을 갖고 있어 문서를 쉽게 HTML 문서로 변환할 수 있다. 그러나 변환된 HTML 문서는 페이지 레이아웃(layout)이 원본과 틀리지며, 그림이 포함된 복합 문서인 경우 그림이 각각의 파일로 구성된다는 문제점이 있다. 또한 변환된 HTML 문서를 재편집하기 위해서는 많은 시간과 비용이 소요되어 효율적이지 못하다.

이러한 문제점을 해결하기 위해 제안된 것이 SGML과 XML이다. 특히 XML은 HTML과 SGML의 장점을 수용하고 단점을 보완한 것으로 관심이 높이고 있다[11]. XML 문서를 만들기 위한 첫번째 작업은 만들고자 하는 문서의 논리적 구조를 표현할 수 있는 DTD를 만드는 것이다. 이
것은 XML 스크립트에서 제시한 문법에 따라 한다. 논리적인 구조만을 가지고 있는 XML 문서가 외부로 보여지기 위해서는 포맷팅 처리가 필요하다. 이를 제공하기 위해 SWG(SGML Working Group)는 SGML의 포맷팅 언어인 DSSSL(Document Style Semantics and Specification Language)을 간소화해서 사용하기로 결정을 했다. DSSSL을 이용해서 포맷팅을 한다면 것은 DSSSL의 스타일 언어를 이용해서 특정 DTD에 대한 스타일 쉘트의 작성은 미한다. 그러나 아직까지 EDMS 시스템에서 XML 문서 포맷을 활용할 수 있는 데는 없으므로 필요에 따라 에플리케이션을 개발하여야 하는 부담이 있다. 하지만 확장성과 활용성이 높은 XML에 대한 연구가 활발히 진행되고 있어 점차 그 이용이 증가할 것이다.

나. PDF 문서로의 변환


PDF 문서로의 포맷 변환은 소스 형태에 따라 중이 문서의 변환과 워드 문서의 변환 두 가지로 구분된다. 소스의 형태가 중이 문서인 경우 우선 스케너로 스캔하여 이미지 포맷으로 변환한 후 다시 PDF 포맷으로 변환하는 과정을 거쳐야 한다. 이 경우 PDF 문서는 텍스트 기반이 아니고 이미지 기반이 되어 PDF의 특성인 텍스트 인식이 되지 않는다는 단점이 있다. 텍스트 기반의 PDF 문서로 변환하기 위해서는 이미지 문서를 OCR 소프트웨어를 이용하여 텍스트로 변환하는 방법이 있으나 본 연구에서는 고려하지 않았다.

원도 기반의 워드프로세서 파일인 경우에는 프린터로 출력하는 것과 같이 PDF Writer를 프린터로 지정하여 출력하여 쉽게 PDF 포맷으로 변환하기가 가능하다. 또한 원도 파일을 포스트스크립트 파일로 변환한 후 PDF 포맷으로 변환하는 방법도 있다. 이미지 파일과 원도 PDF 문서 생성에는 Acrobat이 유용하지만 한글 PDF 문서를 작성하기 위해서는 별도로 개발된 포스트스크립트 드라이버[24]를 이용하여 포스트스크립트 파일로 변환한 후 다시 Acrobat Distiller를 이용해 PDF 문서
로 변환하여야 한다. 또한 한글로 작성된 위드 파일
의 한글 PDF 문서 변환은 'DocuCom'을 이용하
면 쉽게 이루어진다.

다. 전자문서 포맷 변환에 관한 비교

다음은 앞서 제시한 전자문서 포맷들을 이용하
여 EDMS 시스템에서 활용 가능한 문서 포맷으로
의 변환을 실험하고 변환된 문서의 파일크기, 페이지크기 및 해상도 등의 결과에 대하여 비교해 보
았다. 실험은 MS-워드97 위드프로세서로 작성된
문서를 대상으로 5가지 형태로의 문서 포맷 변환을
시도한 결과 (표 1)과 같은 결과를 얻었다.

### 표 1 문서 포맷별 시험 변환 결과

<table>
<thead>
<tr>
<th>문서 포맷</th>
<th>파일크기</th>
<th>페이지크기</th>
<th>비고</th>
</tr>
</thead>
<tbody>
<tr>
<td>위드</td>
<td>1,519KB</td>
<td>23.4KB</td>
<td>MS-Word 97</td>
</tr>
<tr>
<td>HTML</td>
<td>858KB</td>
<td>13.2KB</td>
<td>HTML 파일: 1, 이미지 파일: 21개 (GIF 포맷)</td>
</tr>
<tr>
<td>PDF</td>
<td>891KB</td>
<td>13.7KB</td>
<td>텍스트 인식 가능</td>
</tr>
<tr>
<td>애플리케이션</td>
<td>1,761KB</td>
<td>27.1KB</td>
<td>CCITT G4 Tiff 포맷, 200DPI</td>
</tr>
<tr>
<td></td>
<td>2,730KB</td>
<td>42KB</td>
<td>300DPI</td>
</tr>
<tr>
<td>PDF</td>
<td>1,756KB</td>
<td>27KB</td>
<td>텍스트 인식 불가능, 200DPI</td>
</tr>
<tr>
<td></td>
<td>2,725KB</td>
<td>42KB</td>
<td>300DPI</td>
</tr>
</tbody>
</table>

비교 결과 파일크기와 해상도 부분 등에서 가장
우수한 전자문서 포맷으로 판단된 것은 텍스트 기
반의 PDF 문서 포맷이었다. 즉, PDF 전자문서 포
맷은 위드 파일에서 쉽게 변환 가능하고 원본과
동일한 편집 환경을 유지하면서 하이퍼 링크, 북마
크, 섬네일(thumbnail) 등 다양한 부가 기능을 포함
하여 이용자의 편의성을 높이면서 파일크기가 작아
디스크 공간도 절약할 수 있다는 결과를 얻었다.

### IV. 웹 기반 EDMS의 설계 및 구현

본 연구에서 구현한 웹기반 EDMS는 한국전
력연구원의 연구문서 관리 시스템을 그 대상으로
하였다. 연구문서는 개개 연구원들에 연구 활동 중
발생된 신제품 개발 정보, 아이디어 및 Know-how
등을 작성한 문서를 의미하며, R&D 활동에 필요
한 기술정보로서뿐 아니라 기간 과수의 지적재산
정보로 효율적인 종합관리와 대외적인 보호를 필
요로 하는 정보이다.

1. 시스템 설계

기존 연구문서 관리 시스템의 전문 정보 데이터베이스는 종이 문서 형태의 원문을 약 파일형 시
스템을 이용하여 이미지 포맷으로 변환, 촉적해 왔
다[2]. 즉, 스크너를 이용하여 200 DPI 해상도로 스
캔한 후 Tiff 파일로 저장하였다.

설계 구현한 연구문서 EDMS는 웹 환경으로
전문 정보를 제공하게 되는데 다음과 같은 두 가
지 특징을 갖는다. 첫째 시스템의 외부 환경 변화
에 따른 관련 시스템과 EDMS와의 통합을 추
진하였다. 연구원에에는 전자결재 시스템인 "한호
름"과 업무 전산화를 위한 클라이언트/서버 기술로
개발된 EIS(ETRI Information System)와 연동하
는 연구원 종합정보시스템[5]이 운영되고 있다. 구
현 시스템에서는 EIS와 한호름과의 통합 환경을 구
축하여 연구문서 처리 단계를 줄여 신속하고 효과
적인 연구문서 관리 및 이용 체제를 구축하도록 설
계하였다. 통합 환경으로 다양한 관리 기능이 있는

131
EIS와 전자 결제가 가능한 한호흡을 활용함으로써 연구문서의 관리 기능이 강화되고 연구문서의 시지 및 요약 정보와 원문 파일 전송이 자동적으로 이루어져 연구문서 확득 단계가 신속하게 이루어지게 된다. 즉, 연구부서에서 연구문서가 작성되면 우선 EIS 시스템을 이용하여 문서작성자가 연구문서의 서지사항과 요약을 입력한다. 이렇게 입력된 정보는 EIS 데이터베이스에 수록되어 개인별, 부서별 문서관리 정보로 활용된다. EIS 정보를 한호흡과 연동하기 위하여 연구문서의 원문 파일을 첨부문서로 포함하여 전자결제 기능을 거쳐 결제 연구 문서 등록 절차는 마치게 된다. 결제된 연구문서의 서지 및 요약 정보는 연구문서 EDMS시스템의 입력 정보로 받아들여 종합관리번호가 부여되고, 색인화 추출, 원문 파일 연결 과정을 거쳐 EDMS시스템에서 활용하기 위해 데이터베이스화 과정을 거치게 된다. 연구문서 EDMS시스템과 한호흡 및 EIS 시스템과의 연동으로 연구문서 작성이 되지 않은 이용자 정보활용에 이르기까지 신속한 문서관리체계가 이루어지게 되었으며 또한 EIS를 통한 연구문서 통계 정보와 연구문서 EDMS 시스템의 정보 공유와 전달 기능이 통합되어 종합적인 관리 및 이용 체계가 이루어지게 되었다.

연구문서 EDMS의 두번째 특징은 연구원에서 생성되는 연구문서의 원문 정보에 대한 문서 포맷의 통합을 추진하였다. 전자결제를 위해서 한호흡의 첨부문서로 수신된 연구문서 파일은 EDMS에서 전문 정보를 제공하기 위해서 전자문서 포맷으로 변환하여 데이터베이스로 구축해야 한다. 이 과정에서 구축하고자 하는 원문의 문서 포맷에 대한 통일이 필요하다. 문서 포맷의 통일은 이용자에게 정보 이용의 혼란을 줄이고 편의성을 높여주며 정보를 효율적으로 관리할 수 있게 해준다. 특히 새로운 시스템 통합 환경의 구축으로 기존 MS-Word 파일의 문서 포맷과 호환성이 유지되는 새로운 문서 포맷의 선택이 요구된다. 새로운 문서 포맷으로는 앞서 검토한 바와 같이 기존 문서 포맷과의 호환성을 유지할 수 있고 신속한 원문 정보 데이터베이스 구축이 가능하며 북마크, 삽입기, 하이퍼링크 등 다양한 이용자가 원하는 기능을 포함하고 있는 PDF 포맷을 본 EDMS시스템의 표준 포맷으로 통일하였다. PDF 포맷으로 통일함은 기존 이미지 포맷을 PDF 포맷으로 변환하여야 하는 문제점이 있으나 워드 파일을 이미지 포맷으로 변환하는 것보다는 이용자의 편의성, 정보의 재활용 및 웹 문서 포맷과의 호환성 유지에 바람직하다고 판단하였다.

2. 시스템 구현

가. 입력 시스템

입력 시스템에서는 온라인으로 연구문서 데이터를 입력하고 연구문서의 원문 이미지를 스캐너를 통하여 입력하는 기능으로 구성되어 있다. 온라인으로 입력된 데이터의 색인화과정은 입력 시스템에서 이루어지지 않는다. 그 이유는 업무 흐름이 데이터 입력자와 데이터 관리자가 분리되어 있는 업무 흐름을 따르기 때문이다. 데이터 입력자에 의해 입력된 데이터를 데이터 관리자가 데이터를 점검하고 색인을 편집하는 과정을 거치게 되어 있어 입력 데이터 처리 시간을 단축하기 위하여 색인은 별도의 시스템을 통해하도록 구성하였다.

나. 데이터 로드 시스템

데이터 로드 시스템은 데이터를 일괄 작업(bat-
ch processing)으로 처리할 수 있게 해준다. 특히 기존 시스템에서 새로운 시스템으로 데이터를 이전 하는데 매우 유용하게 구성하였다. 데이터 로드 시스템에서는 많은 데이터를 일괄적으로 처리하기 때문에 데이터의 예를 최소화 할 수 있도록 데이터 예제 체크 기능을 강화하였다. 특히 기존 시스템의 원문 이미지 정보를 업데이트(update)할 수 있는 기능을 포함하고 있다.

다. 색인 시스템

색인 시스템은 연구문서 종류별로 입력된 정보를 관리자나 이용자에 정확히 검색하고 이용하는 색인어를 추출하고 색인 DB(Index Table)을 생성하는 시스템이다. 색인은 연구문서의 도식의 제어 번호를 입력하여 원하는 데이터만을 색인할 수 있도록 구성하였다. 또한 검색 대상 항목도 지정하도록 구성하였다. 색인 시스템은 수행하면 지정한 제어번호가 해당하는 항목의 색인어가 색인되어 새로운 색인 데이터로 구성된다. 제목에 대한 색인은 자동 색인 기법을 도입하였다. 자동 색인은 불용어와 명사 사전을 구축하여 단어를 추출하여 불용어를 제거하고 명사 사전에 의해 색인어를 추출한다. 불용어는 한글과 영문 데이터 1,000개로 구성되었으며, 명사 사전은 59,000 단어로 구성되어 있다.

라. 검색 시스템

검색 시스템은 각 연구문서별로 구성하였으며, 이용자 뿐만 아니라 검색 효율을 높이기 위하여 검색 필드를 구분하였다. 검색 필드간의 검색 연산은 브리언 연산 ‘AND’ 처리하며 검색 필드 내에서는 브리언 연산자(Boolean Operator [AND, OR])를 통해 브리언 연산을 한다. 검색 효율을 높이기 위하여 검색된 전체 결과를 한꺼번에 출력하지 않고 화면 단위로 출력하도록 구성하였다.

검색 단계는 검색어 입력 및 실행, 검색 결과에 대한 간략 정보 보기, 간략 정보 중 원하는 연구문서의 상세 정보 보기 및 원문정보 보기의 4 단계로 구성되어 있다. 원문이 없는 연구문서의 경우 원문보기 버튼이 작동하지 않도록 구성하여 시스템 유효 발생을 막았다. 간략 정보, 상세 정보 보기에서 원하는 정보를 프린터로 출력할 수 있다. 또한, 검색 결과를 제한하여 검색할 수 있으며 검색된 결과 내에서 재검색이 가능하도록 구현하였다. 간략 정보는 텍스트로 구성하여 HTML 포맷으로 제공하며, 전문은 PDF 리더(reader)를 이용해 볼 수 있다.

마. 용어 사전 관리 시스템

용어 사전 관리 시스템은 색인 시스템에서 이용하는 불용어 사전과 명사 사전을 관리하는 시스템이다. 각 사전의 내용을 추가/삭제/변경할 수 있다.

바. 관리/통계 시스템

관리/통계 시스템은 데이터의 입력 현황, 원문 이미지 입력 현황, 기술문서의 등급별 등록 현황 등의 온라인으로 제공하여 효율적인 관리와 별도의 매체를 통하지 않고 연구문서 관리 현황을 조회해 볼 수 있다.

통계 기능은 원서 연구문서 등록 현황, 부서별 등록 현황, 보호 등급별 등록 현황 등이 시계열로 제공된다.

3. 시스템 개발환경

연구문서 종합관리시스템 개발에는 시스템 운영을 위한 전용서버와 원문 정보 구축을 위한 스캔
스테이션 등을 활용하였으며, SYBASE 등의 upgrade를 통해 적정 개발환경을 구축하였다.
- 서비: Hp 9000/800 E45
- DBMS: SYBASE10
- 개발툴: MicroSoft Visual Basic 3.0
- 스캔스테이션: 후지쯔 스캐너
- 기타: ImageMan, C/C++

본 시스템의 하드웨어환경은 (그림 1)과 같다.

(그림 1) 연구문서 관리시스템 하드웨어 환경

4. 시스템 구성도

연구문서 관리시스템의 전체구성도는 (그림 2)와 같다.

5. 시스템의 평가

본 연구에서 구현한 연구문서 EDMS 시스템은 전문 정보를 PDF 포맷으로 변환하여 구축하는 경우 다음과 같은 부분에서 업무수행 활성화 방안을 도출해낼 수 있었다. 첫째, 전문 정보 구축시 소요되는 시간과 비용이 절감되었다는 점이 가장 큰 장점으로 분석되었다. 이는 MS-워드 문서를 HTML로 변환하여 구축하는 경우에 비하여 활동의 신속한 처리가 가능했으며 PDF 포맷으로 변환한 후 제편집함이 필요 없기 때문이다. 둘째, 기존 시스템에서 운영되던 이미지 문서 포맷과 워드 문서를 PDF로 전환하여 인터넷에서 널리 활용하고 있는 전자문서 포맷으로 통일한다는 점이다. 셋째, 연구원 내 그룹웨어와 연계하여 시스템의 통합화를 이루었다는 점이다. 즉, EDMS는 모체기관의 업무에서 사용되는 자료의 저장소로서의 기능을 제공함으로써 워드프로세서, MIS, 워크플로와 같은 시스템간의 통합적인 운영은 대단히 중요하다. 아무리 풍부한 EDMS라 하더라도 특히 MIS 업무와의 비연동은 시스템을 격리시키는 효과를 내고, 시스템의 유용성을 낮추는 요인이 되기 때문이다. 본 연구에서 구현한 연구문서 EDMS는 연구부서에서 작성된 연구문서를 전자결제를 거쳐 자동화되므로 접수하여 데이터베이스로 구축하는 종합적인 관리 체계가 가능하였다. 넷째, 인터넷에서 널리 이용하고 있는 전자문서 포맷을 이용함으로써 문서포맷별로 브라우저를 설치하지 않고 공개 소프트웨어인 Acrobat Reader 또는 DocuReader로 해결이 가능하였다. 이것은 또한 웹 기반의 EDMS의 제한점이기도 하다, 즉 웹 환경의 EDMS는 인터넷 브라우저에서 해결하지 못하는 기능들이 있고 이미지 스케닝이나 편집인 프로그램이 필요하다는 것이 문제점으로 남는다.

본 연구의 또 다른 제한점은 첫째, PDF Reader로 유명한 Acrobat Reader가 비록 공개 소프트웨어이고 널리 이용되고 있지만 이용자에게 부담
지금까지 EDMS에서 전문 정보를 처리할 수 있는 다양한 전자문서 포맷의 특징과 장단점을 비교 분석하여 보았다. 특히, 기존 정보 시스템이 웹 기반으로 전환함에 있어 가장 큰 과제인 정보의 처리 및 구축으로서 보다 효율적이고 경제적인 방법을 제시하고자 하였다.

앞서 서술한 바와 같이 원문 정보는 이미지나 MS-워드 등 다양한 문서 포맷으로 작성된다. 웹 기반의 EDMS에서 이러한 원문 정보를 신속하고, 효율적으로 처리할 수 있는 전자문서 포맷을 선택하는 일은 매우 중요한 부분이 되고 있다.

따라서 본 연구에서는 한국전자통신연구원에서 현재 운용중인 연구문서 EDMS를 사례로 다양한 전자문서 포맷으로 전문 정보를 변환하여 구축한 후 그 장단점을 비교, 분석하였고, 그 결과 변환 처리 기능과 파일 크기 및 해상도 등에서 본 시스템에 가장 적합한 전자문서 포맷으로 판단된 것은 텍스트 기반의 PDF 문서 포맷이었다. 즉, PDF 전자 문서 포맷은 MS-워드 파일에서 쉽게 변환가능하고 원본에 동일한 편집 제한과 하이퍼 링크, 북마크, 스펙일 등 다양한 부가 기능을 포함하여 이용자 편의성을 높이며 파일 크기도 작아 테이터베이스 구축시 디스크 공간을 절약할 수 있다는 결과를 얻었다. 하지만 PDF 포맷이 향후 인터넷의 표준 문서가 될 수 있는 SGML과 XML과의 호환성을 유지하기 위해서는 계속 연구, 발전이 필요하다.
이야 할 것이다. 그리고 보다 완벽한 웹 기반의 연구문서 EDMS를 위해서 전문 검색(Full-Text Retrieval), COLD 및 GUI(Graphic User Interface) 등의 구성요소를 갖춘 기능을 보장할 경우 좀 더 실용적인 시스템으로 발전될 수 있을 것이다.

참 고 문 헌