인터넷환경하에서 효율적 전송을 위한 문서형식에 관한 연구

A Study of Document Format for Effective Transmission on the Internet Environments

조 현 양(Hyun-Yang Cho)*
최 흥 식(Hung-Sik Choi)**

목 차

1. 서 론
2. 전자 문서의 형식
2. 1 저장을 위한 전자문서 형식
2. 2 인터넷을 위한 전자문서 형식
2. 3 기타
3. 전자문서형식 비교
4. 결 론

초 록

최근 급속히 발달한 인터넷을 통해 단순한 전자우편 뿐만 아니라 학술 논문 등 실제 물리적인 문서를 표현하는 전자문서의 교환이 빈번하게 이루어지고 있다. 인터넷 환경에서 문서를 원활히 서비스하고 열람할 수 있도록 하기 위해서는 온라인 문서처리에 대한 해결책이 선행되어야 한다. 특히 이공계 연구개발자들이 생산하는 문헌은 복잡한 수식과 그림, 도표 등을 포함하고 있으며, TeX, 훈글, MS Word 등 다양한 워드프로세서를 사용하고 있다. 이들이 생산한 전자 문헌들을 인터넷에서 온라인으로 제공하기 위해서는 HTML과 같은 전자문서 형식으로의 변환이 우선되어야 한다. 본 연구에서는 현재 사용되고 있는 전자문서 형식들이 가지고 있는 특성과 장단점을 비교 연구하였다. 우선 문서 교환을 목적으로 하는 전자문서 형식이 가져야 할 특성으로 범용성, 신속성, 장치 독립성, 간결성, 확장성 등을 제시하고 이를 기준으로 현재 사용되거나 제안되고 있는 전자문서 형식들을 평가하였다.

ABSTRACTS

Today, we are confronted with huge amount of data which contain complex documents, images and multimedia contents. Therefore a new method is needed to analyze and manage the mathematical expressions and extract new information from them. It is more and more important to manage the document files including mathematical expressions which are generated by general-purpose word processors. Three major word processors are shared over 90% of domestic market. These are HWP, TeX and MS word. Due to the progress of Internet and digital library, it is necessary to develop a system to manage the document file containing mathematical expressions over the Web.

* 연구개발정보센터 문헌정보사업실장
** 서울중앙병원의학도서관 정보관리계 책임사서
접수일자 2000년 3월 4일

1. 서 론

최근 지식정보가 정보화 사회의 핵심역량으로 인식되면서 지식을 기반으로 한 인프라 구축과 동시에 이를 적극적으로 활용하여 새로운 부가 가치를 창출하고 경쟁력을 고양시키는 방안이 모색되고 있다. 특히 통신기술과 컴퓨터 네트워크 기술의 획기적인 발전은 도서관의 업무 처리 및 정보서비스의 양상을 크게 변화시키고 있음은 주지의 사실이다.

인터넷은 통신과 네트워크 기술의 결정체로서 사회 전반에 많은 영향을 미치고 있으며, 도서관의 경우에도 예외는 아니어서 이러한 인터넷 물결에 직접적인 영향을 받고 있다. 따라서 급변하는 디지털 환경에 도서관이 능동적으로 대처하여 보다 나은 서비스를 제공할 필요가 있으며, 이를 위해서 도서관은 보유하고 있는 정보를 컴퓨터와 인터넷을 통해 관리하고 열람이 가능하도록 체계적으로 정리하여 효율적으로 이용자에게 제공할 수 있어야 한다. 이것은 이용자가 물리적 장소인 도서관을 직접 방문하여 필요한 정보를 수집하던 전통적인 도서관과 비교해보면 전혀 새로운 형태의 정보서비스인 것이다. 이러한 역할을 충실히 수행할 수 있는 한 개념으로서 '디지털도서관'이 출현하였고, 미래의 도서관으로 자리매김하고 있다.

디지털도서관은 기존 도서관의 기능 수행은 물론, 사용자로 하여금 시간과 공간에 제약을 받지 않고 이용될 수 있도록 해야 한다. 일반적으로 도서관의 기능은 크게 정보를 수집하여 저장, 관리하는 측면과 이를 검색하여 서비스하는 측면으로 대별할 수 있다. 전자 환경하에서 기존 도서관이 이러한 기본 기능을 수행하면서 인터넷시대에 대비하기 위한 기본 방안은 현재 도서관에서 주종을 이루고 있는 기존 문헌에 대해서는 수서, 대출, 반납 등의 전통적인 관리 방법에다 웹을 통하여 서지정보에 대한 검색과 조회가 가능하도록 연동해 주는 것이다. 이러한 연동 서비스가 부가되면 일단은 인터넷 시대에 대비되었다고 볼 수 있다. 특히 보다 진보된 방안으로서 저작권 문제가 해결되거나 저작권을 보유하고 있는 학위논문이나 보고서와 같은 기존 문서를 스캔하여 원문을 PDF, TIFF 등의 형태로 변환하여 인터넷을 통해 이용자에게 제공할 수 있다면 인터넷 시대에 앞서갈 수 있는 것이다. 물론 이것이 디지털도서관이 가지는 원래의 기능을 충분히 수행할 수 있는 형태는 분명 아니다.

전통적으로 도서관은 단행본과 함께 학술지, 학위논문, 연구보고서, 특허, 기술규격 및 표준, 신문과 잡지 등과 같이 다양하고 수많은 문자정보를 보유하고 있다. 이러한 문자 정보는 대개 문서 혹은 문헌이라는 형태로 담겨져 있는 것이 보통이다. 여기서 문서라 함은 단행본과 같은 인쇄매체를 거친 기존 문서와 인쇄형태를 갖추기 이전 형태로서 컴퓨터 상에 존재하는 전자문서로 나눌 수 있다. 일반적으로 문서의 제작은 대개 워드프로세서로부터 시작된다. 따라서 향후 만들어지는 필사본이나 그림을 제외한 모든 문서는 전자문서로 존재하게 될 것이다. 이러한 전자문서는 기존의 인쇄문헌과 저장 매체가 다를 뿐이지 동일한 내용을 담고 있는 형태이다. 기존의 도서관에서 문자정보를 수록하고 있는 인쇄자료가 핵심으로 다루어진 것과 같이 디지털도서관에서 전자문

서는 중요하게 다루어야 할 사항이다. 특히 인터넷을 통해 정보서비스가 이루어지고 있기 때문에 전자문서의 전송과 관련된 문제는 매우 중요하다.

본 연구는 현재 가장 보편적인 문자정보를 중심으로 도서관에서 인터넷 시대에 능동적으로 대처할 수 있는 방안을 제시하고자 한다. 먼저 인터넷 전자문서의 형식에 대하여 전반적인 내용을 살펴보고, 이를 저장을 위한 전자문서의 형식과 인터넷 서비스를 염두에 둔 전자문서의 형식으로 구분하여 살펴보았다. 또한 각 파일형식의 사이즈를 비교하여 분석하였고, 이에 따른 전자문서 선택에 관한 결론과 제안을 하였다.

2. 전자 문서의 형식

인터넷의 사용이 보편화된 오늘날 과학기술 분야를 중심으로 수식이 포함된 복잡한 문서를 인터넷상에서 직접 사용할 필요성이 제기되고 있으며, 영어권의 경우 이의 효율적인 해결방안으로 JAVA 애플릿이나 CGI, Plug-in을 이용한 갖가지 모델들이 제시되고 있다. 특히 웹서버와 웹 브라우저가 인터넷을 기반으로 활성화되면서 많은 도서관에서도 홈페이지를 구성하여 이용자에게 정보 서비스를 제공하고 있다. 이들 홈페이지에는 특성상 다양한 문서 편집기로 작성된 방대한 양의 논문을 비롯한 문자정보들이 업로드되어 있으며, 이에 대한 검색 요구가 빈번히 발생한다. 그러나 이용자들이 홈페이지를 방문하여 대용량의 논문을 검색할 때 여러 가지 문제점이 노출된다. 즉, 검색속도와 효율성의 문제가 바로 그 것이다. 특히 디지털도서관 체제하에서 웹으로 이용자가 원하는 문헌의 원문을 보려면 전체를 모두 전송 받은 후 화면에 문헌의 첫 페이지가 출력되므로 길이가 길거나 논문의 특정 부분만을 검색하여 화면으로 출력하고자 할 때는 상당한 시간이 요구되어 매우 비효율적이다. 이러한 속도와 비효율성 문제를 개선하기 위해서는 원본 문서 자체의 크기를 최소화하거나 이용자가 원하는 부분만을 보여줄 수 있어야 한다. 또한 디지털도서관 체제하에서의 이용자는 항상 새로운 기술의 편리함을 추구하게 된다. 즉 필요한 정보에 접근하여 이용함에 있어서 불편함이 존재하는 정보는 좀 더 쉽게 구할 수 있는 다른 장소의 비슷한 정보에 비하여 이용이 현저하게 떨어질 수도 있다. 예컨대 인터넷상에서 실제 이용자들이 편리하게 이용할 수 있는 사이트에 몰리게 되는 경우가 좋은 실례라 할 수 있다.

디지털도서관의 핵심은 가능한 한 많은 정보를 디지털화하여 편리하게 이용될 수 있도록 저장하고 있어야 한다. 이러한 작업의 대부분은 자동화되어야 만이 적은 비용으로 소기의 목적을 달성할 수 있다. 특히 문자정보의 디지털화는 전자문서 형식을 사용하여 대부분 이루어지므로 적절한 전자문서 형식의 선택은 디지털도서관의 성패에 결정적인 역할을 한다고 해도 과언이 아니다. 컴퓨터 기술이 하루가 다르게 발전하고 있지만 인쇄 형태의 문헌을 전자 형태로 변환, 입력하기 위해서는 막대한 시간과 예산이 요구된다. 이렇게 막대한 비용과 시간이 투입된 디지털도서관이 활용상의 불편함이나 전송에 따른 속도 문제를 해결하

지 못할 경우 실패로 돌아갈 수도 있다.

전자문서는 크게 세가지 형식으로 나누어진다. 즉 광범위한 공유, 가공 및 사용을 위하여 저장해 두기 위한 형식, 인터넷에서 즉시 서비스가 가능한 형식, 그리고 전자문서를 작성한 워드프로세서의 형식이다. 인터넷상에서 문서의 전송이나 교환을 목적으로 하는 전자문서 형식이 가져야 할 특성으로는 범용성, 신속성, 장치 독립성, 간결성, 확장성 등이 제시되고 있다.

2. 1 저장을 위한 전자문서 형식

저장을 위한 전자문서의 목표는 정보의 공유와 재가공을 용이하게 하는 데 있다. 따라서 이러한 전자문서는 국제 표준을 따르는 것이 일반적이다. 특정 워드프로세서의 형식을 이러한 용도로 사용하는 것도 고려해 볼 수는 있으나 호환성의 문제와 더불어 컴퓨터 소프트웨어가 급격히 진화하기 때문에 추후 사장되거나 사용범위가 줄어들 가능성이 높아 인트라넷과 같은 특수한 환경을 제외하고 디지털도서관에서는 사용하지 않는 것이 바람직하다. 공유와 재가공이 가능한 전자문서는 대체로 내용정보, 구조정보, 양식정보가 결합하여 이루어지는데 TEX, SGML, XML 등이 가장 널리 사용되고 있다. 여기서 재가공이라 함은 예를 들어 학위논문으로부터 그 해의 학위논문 초록 모음집을 따로 만드는 것과 같이 주어진 전자문서 데이터베이스로부터 필요한 정보를 추출하여 새로운 형태의 문서나 통계자료 등을 만드는 것을 의미한다.

(1) TEX

1970년대 말 스탠포드대학교 교수인 Knuth에 의해 만들어진 조판시스템으로서 수식과 도표 처리에 탁월하여 과학기술 문서의 제작에 전세계적으로 널리 사용되어 왔으며, 향후에도 계속 사용될 것으로 전망된다. 1980년대 초 Lampert에 의해 만들어진 매크로 패키지인 LATEX은 TEX의 강력한 조판 컴파일러의 기능을 충분히 활용하여 TEX을 단순한 조판 기능을 가진 워드프로세서에서 문서내용을 구조정보와 함께 입력하여 양식정보가 담긴 스타일파일과 함께 처리하여 원하는 출력을 얻는 오늘날의 저장을 위한 전자문서 형식으로 발전시켰다.(Wilson, 1998) LATEX과 결합된 TEX은 목차와 상호참조를 자동으로 갱신하는 기능 뿐 아니라, 색인과 참고문헌도 자동으로 처리하는 현재의 워드프로세서도 가지고 있지 못한 기능들을 추가함으로써 전자문서 표준규격인 SGML의 제정에 지대한 영향을 주었다.

한편 TEX은 멀티미디어 관련 기술이 폭발적으로 발전되기 전에 개발되었기 때문에 TEX에는 이들을 처리하는 규격은 정해져 있지 않다. 그러나 현재 대부분을 차지하는 문자정보와 삽입된 그림 위주의 문서에서는 아직도 강력한 도구로서 사용되고 있다. 또한 미국을 비롯한 서양 언어권에서는 대부분의 자연과학 출판사가 조판시스템으로 사용하고 있고, 저자에게 TEX으로 원고를 작성할 것을 권장하는 학술지도 많이 있다. LATEX의 강점은 위에서 살펴 본 바와 같이 강력한 조판 컴파일러 기능을 가지고 있으며, LATEX으로 작성된 원고로부터 SGML로의 자동 변환이 가능하다. 실제로 이와 같이 SGML로 보관하여

향후 어떤 변화에 대비하는 출판사도 있다.

결론적으로 TEX/LATEX의 조합은 문자정보 위주의 문서에서는 저작도구로써 뿐만 아니라 저장을 위한 전자문서의 형식으로도 향후의 변화에 별 다른 어려움 없이 적용이 가능하다. 또한 인터넷에서 제공 가능한 전자문서인 DVI나 PDF로 실시간에 추가 비용없이 만들어 제공할 수 있다는 장점도 있다.

(2) SGML

SGML은 1980년 중반 국제표준기구(ISO)가 제안한 국제문서규격으로 문서의 구조정보를 기술하기 위한 규약을 정한 메타언어이다.(ISO/IEC JTC1/SC34,1998) SGML(Standard Generalized Markup Language)로 전자문서를 만들려면 문서의 논리구조를 따로 정의한 DTD(Document Type Definition)가 먼저 만들어지는 것이 보통이며, 이 구조에 맞는 내용을 입력하게 된다. SGML문서로부터의 출력은 좀 더 복잡하다. DTD에 등재된 문서구조에 따라 원하는 모양으로 출력하는 양식을 작성하여 출력프로그램을 작성하는 것이 한가지 방법이나, 이것은 구현하는 방법에 너무 많이 의존하게 되어 정보의 광범위한 공유라는 원래의 취지에 벗어날 수도 있다. 이를 극복할 수 있는 한가지 방법은 DSSSL(Document Style Semantics and Specification Language)이라는 또 하나의 SGML 문서를 변환하여 조판이 가능한 다른 SGML문서로 만들어 주는 국제표준규격을 사용하는 것이다. 이 경우 DTD는 DSSSL규격에 맞게 작성되어 있어야 할 필요가 있다. 실제로 SGML/DSSSL의 조합으로 작성된 전자문서는 추가 비용이 없이 TEX이나 Microsoft Word로 원하는 출력물을 얻을 수 있다. 또한 TEX/LATEX을 사용하여 실시간에 DVI나 PDF로 만들어 인터넷에 제공하는 것도 가능하다. DSSSL 규격을 따르지 않은 SGML문서를 인터넷에 제공하려면 이를 해석하는 특수한 프로그램이 있어야 하고 이는 비용측면 뿐 아니라 효율성 측면에서도 현명하지 못한 선택일 수 있다.

결론적으로 SGML은 신중한 DTD의 작성이 전제된다면 문자와 멀티미디어를 포괄하는 정보의 저장을 위한 전자문서로서 바람직하다. 그러나 다양한 워드프로세서 형식의 문서를 SGML로 만들기 위해서는 현재 막대한 시간과 비용을 감수해야 하고, 규약의 지나친 일반성 때문에 국제무역거래 등의 특정 목적을 제외하고는 널리 사용되지 않고 있다는 점을 간과해서는 안된다.

(3) XML

다양한 DTD에 의존한 SGML문서를 모두 웹 브라우저에서 출력하도록 만드는 일은 사실상 불가능하다. 설령 가능하다고 해도 SGML문서가 네트웍의 대역을 너무 많이 차지하게 되어 효율이 떨어지게 된다. 그리고 현재 웹 전자문서의 표준인 HTML(HyperText Markup Language)은 문법구조 확장이 한계에 도달하여 복잡한 수식이나 그림을 포함하고 있는 문서는 HTML로 변환해서는 원래의 형식을 살리기가 사실상 불가능한 것처럼 보인다. 따라서 이 두 형식의 장점만을 취하여 웹에서 제공하는 것을 궁극적인 목표로 W3C(World Wide Web Consortium)에서 제정한 전자문서 표준형식이 XML(Extensible

Markup Language)이다. XML문서도 미리 작성된 DTD를 따라 입력하는 것이 보통이다. 그러나 XML의 DTD는 SGML의 DTD와 비교하여 단순화되고 불필요한 사양들이 생략될 수 있다는 것이 다르다. 따라서 XML은 SGML의 부분집합이라고 생각해도 무방하다.

XML에서도 문서의 출력을 구현하는 것이 가장 어려운 문제이다. 현재로서는 SGML의 DSSSL과 넷스케이프사의 HTML을 위한 출력 양식인 CSS(Cascading Stylesheet Specification)의 절충형인 XSL(Extensible Style Language)이 제안되고 있으나 XSL에 관한 표준은 아직 정해지지 않고 각 회사마다 규격이 다른 것이 문제로 남아 있다. 반면 XML은 향후 W3C 표준인 MathML을 포함할 예정이어서 수식의 처리에도 어려움이 없을 것이다. 또한 레이어의 동작을 제어하는 Dynamic HTML의 기능을 수행한 DOM(Document Object Model)도 아울러 표준화되고 있고, 네트웍과 컴퓨터의 성능이 급격히 향상되는 점을 감안하면 가까운 장래에 저장을 위한 전자문서 뿐 아니라 인터넷을 위한 전자문서의 가장 강력한 표준으로 자리잡을 가능성이 높다. 현재로서도 XML/XSL의 조합으로 작성된 전자문서는 XSL을 변환과정과 TEX 컴파일러를 이용하여 실시간에 DVI 또는 PDF로 만들어 추가 비용 없이 인터넷에 제공하는 것이 가능하다.

결론적으로 XML은 정보의 공유와 재가공이 목적인 디지털도서관에서 채택할 수 있는 전자문서 형식 가운데 하나이다. 그러나 XSL의 표준이 명확히 정의되지 아니한 상황이어서 표준이 확정될 때까지는 일단 관망하는 것도 또한 현명한 방법이다. 더욱이 특정 워드프로세서 형식을 XML로 변환해 주는 도구를 해당 워드프로세서 제작사가 만들 때까지는 변환시 추가 비용을 감수해야 하는 문제를 지니고 있다.

2. 2 인터넷을 위한 전자문서 형식

인터넷서비스를 위한 전자문서의 목표는 서비스의 효율과 문서의 질을 높이는 데 있다. 특정 워드프로세서 형식을 이 목적으로 사용하는 것은 일단 파일이 인터넷에서 서비스하기에는 너무 크고, 받는 대로 보여주는 스트림(Stream) 방식의 처리가 어렵다. 따라서 이러한 전자문서는 정보의 빠른 획득을 원하는 사용자들이 선호하지 않는다. 다만 연구계획서의 작성과 같이 특정 워드프로세서를 보유하고 있는 사용자가 인터넷에서 그 워드프로세서로 만들어진 연구계획서 양식을 받아 재사용 할 경우는 예외일 수 있다. 이 경우 연구계획서 양식은 인터넷을 위한 전자문서의 개념이기보다는 일종의 다운로드 받는 자료라고 해야 옳다. 그리고 디지털도서관에서 저자가 작성한 워드프로세서 파일을 그대로 제공하는 것은 정보의 무분별한 재사용을 초래하여 저작권에 문제가 있을 수도 있다는 사실을 간과해서는 안될 것이다.

인터넷을 위한 전자문서가 소기의 목적을 달성하기 위하여 무엇보다 중요한 것은 먼저 파일의 크기를 줄이고 플랫폼의 독립이 이루어져야 한다. 따라서 내용, 글꼴정보, 위치, 삽입된 기타 오브젝트의 정보만을 전자문서에 담는 것이 보통이다. 이러한 전자문서 형식으

로는 PDF, DVI 등이 있다.

전자문서에 포함되는 내용 가운데 글꼴정보는 파일의 크기에 지대한 영향을 미치므로 대부분의 글꼴은 정보의 크기가 큰 글꼴 모양은 담지 않고 글꼴명과 크기, 굵기 등의 간단한 정보만 담고 보여 줄 때는 사용 가능한 글꼴 중 가장 가까운 것으로 나타나게 된다. 특히 글자 수가 많은 한글과 한자의 경우 글꼴 모양 정보의 분량은 상당히 크다. 이 점이 대량의 전자문서를 보유할 디지털도서관에서는 가장 유의해야 할 사항이다. 왜냐하면 각 전자문서마다 중복해서 들어가 있는 글꼴 모양은 저장장치에 상당한 부담을 줄 수 있을 뿐만 아니라 결과적으로 인터넷에서 제공 속도에도 크게 영향을 미친다. 따라서 한글 인터넷 전자문서는 글꼴의 모양을 담지 않는 것이 훨씬 유리하고 고정된 폭이 대부분인 한글에서는 글꼴을 대체하는 것이 크게 문제되지 않는다.

(1) PDF

PDF는 90년대 초 미국의 아도비(Adobe)사가 만든 전자문서 형식이다. 현재는 웹에서 제공하는 전자문서의 형식으로 서양 언어권에서는 많이 채택되고 있지만, 원래의 목적은 DTP(Desk Top Publishing)을 위해 문서를 상호 교환이 가능하도록 하기 위한 것이었다. DTP를 위해서는 그래픽 패키지를 포함한 많은 응용프로그램과 다양한 글꼴이 사용되기 마련이었고 이러한 문서는 생성한 모든 응용프로그램과 글꼴이 갖추어지지 않은 시스템과 교환하거나 외부 출력소에서 인쇄하는 데 어려움이 있었다. 이러한 어려움을 극복하기 위해 프린터 구동을 위한 아도비 사의 소유 형식인 포스트스크립트를 압축하고 열람의 편의를 위해 하이퍼 텍스트 등의 첨가가 가능한 형식으로 발전시킨 것이 PDF형식이다.

PDF형식은 아도비 사의 이해관계 상 그들의 포스트스크립트 글꼴형식 만을 사용한다. 바로 이 점이 하나의 글꼴 파일의 크기가 엄청난 2바이트 글자 언어권에서는 PDF형식의 문서가 사용되기가 어려운 이유이다. 마이크로소프트 윈도우즈와 맥킨토시 OS에서는 아도비 글꼴형식에 대항하여 트루타입(True Type)이라는 외각선 글꼴 형식이 사용되어 왔다. 따라서 이러한 OS 상에서 PDF문서를 처리하기 위해서는 글꼴을 PDF문서에 내장하거나 사용된 포스트스크립 글꼴을 미리 사용자가 컴퓨터에 가지고 있어야 한다. 사용자가 포스트스크립 글꼴을 컴퓨터에 가져야 할 경우 글꼴 파일의 크기가 비교적 작은 서양 언어권에서는 문제가 되지 않는다. 그러나 글꼴 파일의 크기가 큰 2바이트 글자 언어권에서는 글꼴이 PDF기술에 커다란 부담으로 작용하여 특히 웹에 제공하는 전자문서로서의 기능에 장애물이 되어 왔다. 그리고 또 다른 문제는 웹 브라우저와 연동시 필수 기능이라고 할 수 있는 스트림 처리가 PDF문서의 특성상 곤란하다는 것이다. 더욱이 아도비사 자신도 XML 컨소시움에서 주도적으로 활동하고 있어 인터넷 전자문서로서 PDF 형식을 고집하지는 않을 전망이다. 다만 PDF는 원래의 목적인 출판용으로는 계속 사용될 공산이 크다.

(2) DVI

인터넷을 통한 문서 전송에 있어서 필수적인 사항은 문서의 간결성은 물론 온라인으로

제공하기 위해서는 HTML과 같은 전자문서 형식으로의 변환이 우선되어야 한다. 가장 단순한 전자문서 형식으로 문자집합을 순서대로 나열한 텍스트 형식이 있다. 텍스트 형식은 특정한 인코딩 방식에 동의가 이루어져 있는 한 어떠한 컴퓨터에서도 동일한 내용을 표현하게 된다. DVI는 문서의 내용을 표시하는데 있어 최소한의 정보만을 수록하여 대단히 간결하다는 것이 가장 큰 장점이다.

이러한 텍스트 형식은 실제로 인터넷을 통한 파일의 전송이나 전자우편 등에서 가장 빈번히 사용되고 있다. 그러나 우리가 실제 사용하는 여러 종류의 문서가 텍스트 형식과 같이 문자 나열 정보만을 가지는 경우는 거의 없다. 문서에서 사용하는 글꼴에 관한 정보 및 문자가 문서 내에 표현될 위치의 좌표에 관한 정보가 최소한으로 포함되어 있어야 한다. DVI 형식은 글꼴 정보 및 문자 위치정보를 최소한의 용량만 사용하여 표현하는데 주안점을 둔 형식이라 할 수 있다.

MS Word나 훈글 등 대부분의 상용 문서편집기는 각각 전용으로 사용하는 전자문서 형식을 가지고 있다. 이들 전자문서들 역시 글꼴 및 문서구조 정보를 포함하여야 하므로 DVI 형식이 가지는 정보들을 포함하고 있다. 그러나, 이들 형식은 문서의 편집을 용이하게 하기 위하여 많은 종류의 부대적인 정보를 수록하고 있기 때문에 문서의 용량이 비효율적으로 커지게 된다. 즉 문서를 편집하는데 목적을 두지 않고 오직 문서를 표현하고 전달하는 것에만 목적을 둔다면 전자문서 형식은 훨씬 간결한 형식만으로도 충분하다. DVI는 이러한 목적에 잘 부합되도록 설계되었다.

많은 경우에서 문서는 글꼴 및 문자만으로 표현할 수 없는 정보를 포함하고 있다. 전통적인 문서의 경우에는 사진, 그림 등 화상정보를 예로 들 수 있으며, 최근의 전자문서는 하이퍼링크, 음성정보 등이 포함되는 경우가 빈번하다. 이들 정보는 그 특성상 장치 독립적이기가 어렵고 문자 집합과는 달리 문서의 중심 내용이 되기는 어려우며 관련된 표준안이 계속해서 변화하고있기 때문에 DVI 형식에 삽입하는 것은 자연스럽지 못하다. 그러나, DVI 형식은 장치 비독립적인 정보를 수용하기 위한 방법 또한 제공하므로 위와 같은 정보 또한 수록할 수 있다.

DVI 형식 내에서 가장 빈번히 사용되는 명령들은 문자 표시 명령, 띄어쓰기에 해당하는 좌표 이동 명령, 줄 바꿈에 해당하는 좌표 이동 명령, 글꼴 변경 명령, 좌표의 스택에 관련된 명령 등이 있다. DVI내에서 이들 명령은 모두 1 바이트만을 사용하도록 구현되어 있다. 문서를 표현하기 위해서 가장 최소한으로 필요한 정보는 문자 집합과 사용 글꼴, 그리고 각 문자의 문서 내 위치 좌표이다. 각 문자의 좌표를 표시하기 위해 각 문자 표시 명령을 쓸 때마다 좌표를 함께 표시해 주는 방법을 고려할 수 있다. 그러나 이러한 방법은 현재 일상적으로 사용하는 문서의 형태를 고려할 때 대단히 비효율적이다. 일반적으로 문서 내에서 한 문자를 표시하고 나면 그 다음 문자는 가로 또는 세로 방향으로 일정한 양만큼을 이동하여 표시된다. 즉 문자를 표시할 절대 좌표보다는 앞 문자와의 상대 좌표를 사용하는 것이 효율적이며, 띄어쓰기나 줄 바꿈 간격이 일정할 때는 매번 그 간격을 기술하기보다는

1 바이트만을 차지하는 명령을 사용하는 것이 효율적이다. 이와 같이 DVI 형식은 현실에서 일상적으로 사용되는 문서들의 특징을 고려하여 최소한의 용량만으로 문서를 표현할 수 있도록 설계되었다.

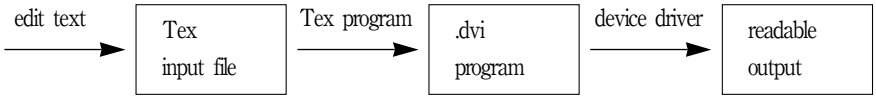
DVI 형식에서 1 바이트 이상을 사용하는 명령에는 글꼴 정의 명령, 페이지 시작 명령, 프리앰블, 포스트앰블 등이 있다. 글꼴 정의 명령은 글꼴의 이름과 글꼴을 불러들일 때 사용할 확대 비율 및 DVI 문서 내에서 해당 글꼴을 불러올 때 사용하는 글꼴번호로 이루어져 있다. 페이지 시작 명령과 포스트앰블에는 특히 문서내의 페이지 단위 검색을 빠르게 하기 위하여 직전 페이지 시작 명령의 위치가 기록되어 있다. 이와 같이 DVI 형식의 전자문서는 물리적인 문서로 출력할 수 있을 뿐 아니라 여타 다른 매체를 통하여 효율적인 열람 및 검색이 가능하다.

DVI 형식은 위에서 전자문서 형식이 갖추어야 할 조건으로 열거한 간결성, 장치독립성, 범용성, 보편성의 측면을 모두 갖추고 있어서 디지털도서관에서 사용할 자료형식으로 가장 적합하다고 할 수 있다. 이미 설명된 바와 같이 DVI 형식은 문서에 대한 정보를 다른 어떤 문서 형식보다 효율적으로 문서를 기술함으로써 간결성을 보장한다. 또한 DVI는 "DeVice Independent"의 약어로 DVI 형식이 설계된 이유 자체가 장치독립성을 보장하는 전자문서 형식의 필요성 때문이어서 장치 독립적이고, 1980년대에 개발되고 난 이후 과학기술분야를 포함하여 전 분야의 문서형식으로 광범위하게 활용되어 그 범용성과 보편성이 검증되었다. 따라서 본 연구는 DVI 형식을 현재 디지털도서관을 구축하는데 사용할 수 있는 가장 적합한 전자문서 형식으로 제안한다.

이러한 DVI 파일이 TeX에서 생성되는 절차는 우선 출력할 내용을 담고 있는 입력 파일을 작성하여 이 파일을 에디터를 이용하여 ASCII 형태로 작성한 다음 TeX 프로그램으로 작성한 입력 파일을 컴파일하여 .dvi 파일을 생성한 후에 이 DVI 파일을 읽고 원하는 형태로 출력해 주는 출력장치 프로그램을 이용하여 출력하면 된다. 이와 같은 절차를 그림으로 나타내면 〈그림 1〉과 같다.

이렇게 출력을 얻기까지 약간은 복잡한 절차가 있지만 이렇게 함으로써 사용자에게는 작성한 입력 파일이 ASCII 형식이므로 어느 컴퓨터에서도 작업할 수 있고, 또 출력이 .dvi 파일로 나오므로 이 파일을 가지고 어떠한 출력기기에서도 출력할 수 있다는 편리한 점이 생긴다.(박기현, 김철수, 1995)

이와 같이 DVI는 TEX 문서를 지정된 양식과 함께 TEX 컴파일러를 통과하여 얻는 전자문서 형식으로서 내부형식이 PDF에 비해 엄밀하지만 간결하다. 이러한 DVI로 작성된 문서는 이진(Binary) 형식으로 매우 효율적으로



〈그림 1〉 TeX에서 DVI 파일이 변환되는 과정

쓰여 삽입된 그림이 없는 경우 압축해도 절반 이하로는 줄지 않는다. 또한 DVI는 컴퓨터의 자원이 지금처럼 풍부하지 못하던 시절에 만들어져 저장공간을 최소화하면서 처리속도를 빠르게 하려는 지혜가 돋보인다. 그리고 DVI는 글꼴 정보만을 담고 모양자체는 포함하지 않기 때문에 파일의 크기를 최소화할 수 있다. 따라서 인터넷에 제공하는 전자문서로서 적절하다고 하겠다. 그 동안 DVI의 이러한 장점에도 불구하고 인터넷을 위한 전자문서 형식으로 사용되지 못한 것은 웹 브라우저와 연동하는 출력기가 없었기 때문이다. 위에서 언급한 바와 같이 SGML/DSSSL 문서나 XML/XSL 문서가 실시간에 TEX 컴파일러를 통하여 DVI로 변환될 수 있어 한글을 포함하고 있는 인터넷 전자문서로는 가장 적절한 형식이라고 할 수 있다.

2. 3 기타

현재 웹의 표준형식인 HTML을 생각해 볼 수 있다. HTML(Hypertext Markup Language)은 WWW상에서 텍스트 및 관련데이터의 교환을 위한 문헌의 형식으로서 플랫폼에 독립적인 하이퍼텍스트 문헌을 작성하는데 이용되는 단순한 마크업 언어이다. HTML 문헌은 다양한 영역의 정보를 표현하는데 적절한 범용 의미를 가진 SGML 문헌이다. HTML 마크업은 하이퍼텍스트 뉴스, 메일, 도큐멘테이션, 하이퍼미디어, 옵션 메뉴, 데이터베이스 질의 결과, 삽입 그래픽이 있는 간략 구조화 문헌, 그리고 하이퍼텍스트형 표현이 가능하다.

HTML은 1990년 이래로 널리 이용되고 있다. 버전 2.0(RFC 1866)은 1994년 이전에 널리 이용되었던 HTML의 기능과 대체로 상응한다. 1997년에는 HTML 4.0버전이 발표되어 1997년 12월 18차 W3C 권고안으로 승인되었다. 확장된 기능은 다중언어 데이터 표현, 상호작용 요소 및 객체 그리고 캐스캐이딩 형식을 이용한 표현의 통제 등이 포함되어 있다.

ISO는 안정된 분산 플랫폼이 요구되는 문서의 작성 이용을 잘 지원하는 HTML 태그세트 규정하는 표준초안 ISO/IEC FCD 15445: Information technology -- Hypertext Markup Language (ISO-HTML)를 제시하였다. 그러나 이러한 HTML은 복잡한 전자문서에 대한 표현 및 향후 확장성에 있어서 한계에 도달해 있으므로 신문, 잡지와 같이 단순한 문서가 아니고는 고려 대상에서 제외된다.

그 밖에도 포스트스크립트 또는 레이저 젯 프린터로 보내지는 형식인 PS 또는 PCL이 있다. Postscript는 1985년 미국 Adobe사에서 개발한 인쇄된 페이지의 형식을 기술하는 프로그래밍 언어로서 그 주된 목적은 장치에 관계없이 똑같은 이미지를 나타내기 위한 언어를 제공하고자 함에 있다. Postscript는 벡터 및 래스터그래픽을 호환 가능한 프린터에서 처리에 적합한 형태로 통합한 포맷된 텍스트파일로서 ① 강력한 그래픽기능을 갖는 범용 프로그래밍언어 ② 프로그래밍기능을 갖고 있는 페이지기술언어 ③ 래스터출력장치를 통제하기 위한 대화형 시스템(interactive system)이라는 특징을 갖는다. 임의의 Postscript 파일은 파일의 수신자 입장에서는 편집이 불가능하다. 따라서 이용자가 파일을 쉽게 변경시킬 수 없으므로 외형을 원형대로 보존할 수 있고 저작권의

보호가 용이하다는 장점이 있다. 그러나 Postscript 문서에서 사용된 폰트가 이용자의 컴퓨터에 없을 경우에는 문제가 발생하게 되며, 특히 특수 폰트를 사용하는 문헌의 경우에는 컴퓨터에 기본적으로 내장된 폰트로는 원형에 가까운 문서를 출력할 수 없게 된다는 점과 전자문서이기에는 글꼴을 비효율적으로 다루어 파일의 크기가 매우 커지는 단점이 있다.

3. 전자문서형식 비교

오늘날 소유에서 접근으로의 패러다임이 변화함에 따라 서지 데이터베이스보다는 전문 데이터베이스에 대한 요구가 더 많고, 더 나아가서는 텍스트, 음성, 그래픽, 그림의 정지화상 또는 동화상을 포함한 멀티미디어 데이터베이스에 대한 선호도가 날로 증가하고 있다. 따라서 도서관은 이러한 변화와 정보요구를 적극 수용하기 위하여 인쇄자료에서 멀티미디어 자료에 이르기까지 다양하고 많은 자료를 수집해야 하고, 신속한 정보서비스가 이루어 질 수 있도록 잘 관리되어 있어야 한다. 현재 많은 도서관에서 이를 실현하기 위하여 기존 자료를 디지털화하고, 각종 파일형태의 자료를 수집하여 저장하고 서비스하는 디지털도서관을 구축하는데 많은 관심과 노력을 기울이고 있다.

본 연구는 인터넷상에서 어떤 전자문서형식이 효율적인 전송을 할 수 있는지를 실험하기 위하여 전송에 많은 영향을 미치는 파일 사이즈를 비교하여 보았다. 즉 각종 워드프로세서에서 작성된 문서를 디지털도서관 구축에 사용할 수 있는 여러 가지 포맷으로 변환하여 그 파일 크기를 비교하였다. 여기에서 파일사이즈를 비교한 이유는 많은 정보를 수집하여 저장하고, 이를 이용자에게 빠른 속도로 전송하기 위해서는 파일의 크기가 매우 중요하기 때문이다. 물론 각종 파일 포맷과 관련하여 국제적인 표준화 동향을 참고하여 신중을 기해야 함은 당연하다.

본고에서 실험한 환경은 동일한 내용을 각각의 편집기로 작성하여 이를 A4 1장에 수록되도록 하여 실험하였다. 실험에 사용한 워드프로세서는 현재 국내에서 많이 사용하고 있는 글, 한글워드, 훈민정음, 엑셀, 파워포인트를 대상으로 하였고, TeX은 자체적으로 DVI 파일을 작성할 수 있기 때문에 제외하였다. 이를 비교한 결과는 〈표 1〉과 같다.

〈표 1〉은 현재 국내에서 문서 작성시 많이 사용하고 있는 각종 워드프로세서를 이용하여

〈표 1〉 파일 사이즈 비교표(단위 : KB)

문서형식	원파일	PDF	TIFF	DVI	압축DVI	비고
아래한글	27(100%)	65(241%)	47(174%)	10(37%)	4(15%)	
한글워드	24(100)	11(46)	46(192)	10(42)	4(17)	
훈민정음	42(100)	11(26)	46(109)	10(24)	4(10)	
엑셀	17(100)	4(24)	38(224)	10(59)	4(24)	
파워포인트	48(100)	12(25)	55(115)	10(21)	4(8)	
평균	31.6(100)	20.6(65.2)	46.4(146.8)	10(31.6)	4(12.7)	

디지털도서관에서 구축할 수 있는 다양한 형태의 포맷으로 변환시켜 본 결과이다. 평균적으로 31.6KB인 원파일(source file)을 각각의 다른 포맷으로 변환한 결과, 압축DVI파일이 4KB로 가장 적게 나타났고, DVI가 10KB, PDF가 20.6, TIFF가 46.4로 나타나 압축 DVI가 저장과 전송의 측면에서 가장 효율적인 전자문헌 형식으로 나타났다.

이를 좀더 상세하게 파일크기와 백분율로 구분하여 살펴보면 다음과 같다.

첫째, 파일 사이즈만을 보았을 경우 원파일의 크기는 파워포인트(48KB)로 작성된 경우가 가장 크게 나타났고, 엑셀(17KB)이 가장 적게 나타났으며, 이를 PDF 형식으로 변환한 결과는 훈글파일을 PDF(65KB)로 변환한 경우가 가장 크게 나타났고, 엑셀파일을 PDF(4KB)로 변환한 경우가 가장 적게 나타났다. 또한 이를 FIFF로 변환한 결과는 파워포인트 파일을 TIFF(55KB)로 변환한 경우가 가장 크게 나타났고, 엑셀파일을 TIFF(38KB)로 변환한 경우가 가장 적게 나타났으며, DVI와 압축 DVI로 변환한 결과는 모두 10KB, 4KB로서 동일한 크기로 변환되었다.

둘째, 원파일(100%)에 대한 백분율로 살펴보면 PDF로 변환한 결과는 훈글파일을 PDF(241%)로 변환한 경우가 가장 크게, 엑셀파일을 PDF(24%)로 변환한 경우가 가장 작은 것으로 나타났다. 또한, 원파일을 TIFF파일로 변환한 결과를 살펴보면 엑셀파일을 TIFF(224%)로 변환한 경우가 가장 크게, 훈민정음 파일을 TIFF(109%)로 변환한 경우가 가장 작은 것으로 나타났다. 마지막으로 원파일을 다시 DVI 및 압축 DVI로 변환하였을 때 엑셀 파일을 DVI(59%), 압축 DVI(24%)로 변환한 경우가 가장 크게 나타났고, 파워포인트를 DVI(21%), 압축DVI(8%)로 변환한 경우가 가장 작게 나타났다.

결과적으로 PDF파일은 4-65KB로 변환된 파일 사이즈의 범위가 크게 나타났고, FIFF는 대체적으로 38-55KB로 비슷하게 나타났으며, DVI(10KB)와 압축 DVI(4KB)는 동일한 사이즈로 나타났다. 파일의 크기는 인터넷 환경 하에서 전송에 따른 속도를 극대화 하는데 절대적 영향을 미치는 것을 고려해 볼 때 디지털도서관 구축에 있어서 DVI형식은 저장과 전송의 측면에서 최고의 방안으로 선택될 수 있을 것이다.

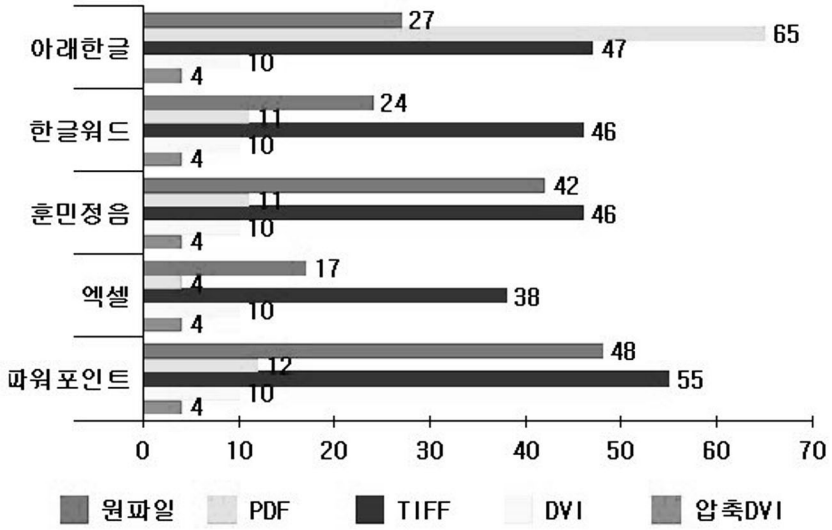
이상에서 살펴본 파일사이즈의 변화를 그래프로 나타내면 〈그림 2〉와 같다.

4. 결 론

본 연구에서는 인터넷상에서 문서 교환을 목적으로 하는 전자문서 형식이 가져야 할 특성으로 범용성, 신속성, 장치 독립성, 간결성, 확장성 등을 제시하고 이를 기준으로 현재 사용되거나 제안되고 있는 전자문서 형식들을 평가하였다.

전자문서형식을 조사하고 실험한 결과는 다음과 같다.

첫째, 각종 문서편집기에서 작성된 원파일의 평균 사이즈가 31.6KB인 파일을 각각의 다른 포맷으로 변환한 결과, 압축 DVI파일이 4KB로 가장 작은 것으로, DVI가 10KB, PDF가 20.6, TIFF가 46.4로 나타나 압축 DVI가 저



〈그림 2〉 파일 사이즈 비교 그래프

장과 전송의 측면에서 가장 효율적인 전자문헌 형식으로 나타났다.

둘째, 원파일을 다른 파일형식으로 변환된 파일사이즈의 범위는 PDF파일이 4-65KB로 가장 크게, TIFF는 대체적으로 38-55KB로 비슷하게 나타났으며, DVI(10KB)와 압축 DVI(4KB)는 동일한 사이즈로 나타났다. PDF가 넓은 범위 내에서 변환된 이유는 혼글에서 큰 차이를 보였기 때문이다. 즉 혼글의 원파일이 27KB(100%)인 것이 PDF로 변환되면서 65KB(241%)로 파일사이즈가 약 2.5배 커진 것이다. 이것은 국내에서 많이 사용되고 있는 문서편집기가 혼글이라는 점을 감안한다면 전자문서형식 선정에 신중을 기할 필요가 있다.

DVI는 TEX 문서를 지정된 양식과 함께 TEX 컴파일러를 통과하여 얻는 전자문서 형식으로서 내부형식이 PDF에 비해 엄밀하지만 간결하다. 또한 DVI형식이 파일 크기에 있어서도 가장 효율적으로 변환되어 여타 전자문서 형식에 비하여 인터넷 환경 하에서 문서의 전송 및 저장에 가장 적합한 전자문서형식으로 나타났다.

참 고 문 헌

박기현, 김철수. 1995. TeX 입문. 서울. 경문사. p.8.

"Introduction to XSL", 〈http://www.sil.org/sgml/x니.html#intro,〉 [1998년 7월 2일]

Welcome to the ISO/IEC JTC1/SC34 Web Service. 〈http://www.oml.gov/sgml/sc34home.htm.〉 [1998년 6월 10일]

Wilson, PR. "LTX2X: A LaTeX to X Auto - tagger", 〈http://ctan.unsw.edu/tex-archive/support/ltx2x/ltx2x.html〉 [1998년 7월 3일]

〈http://www.adobe.com〉 [2000년 3월 6일]

〈http://www.oasis-open.org〉 [1999년 12월 22일]

〈http://www.w3.org〉 [1999년 12월 28일]

〈http://www.tug.org〉 [1999년 12월 22일]