원문이미지 구축에 있어서의 데자뷰(DjVu) 포맷의 적용사례

서울대학교 도서관
김 성 중

< 목 차 >

1. 서언
2. 데자뷰 포맷의 도입 배경
3. 데자뷰 포맷
4. 데자뷰 포맷 서울대학교 적용 사례
5. 결 론

1. 서 언

대학이나 주요 기관들이 원문컨텐츠를 구축하여 서비스하는 전자도서관을 추진하고 있거나 추진할 계획을 세우고 있다. 서울대학교에서는 원문 컨텐츠 구축을 캠퍼스 내에 있는 전 기관을 대상으로 하면서 다양한 유형의 자료가 컨텐츠 구축의 대상이 되고 있다. 다양한 유형의 자료를 고품질로 컨텐츠를 구축하다 보니 원문 서비스 포맷의 한계가 있었다.

이러한 한계를 극복하기 위하여 원문 컨텐츠 중의 서비스 포맷의 일부를 데자뷰(DjVu) 포맷으로 국내 최초로 적용하였다. 이러한 데자뷰 포맷의 소개, 도입배경, 적용대상, 서울대학교의 적용 사례를 소개함으로써 전자도서관을 추진하는 도서관에게 도움이 되고자 한다. 원문 서비스에 있어서 PDF가 한계가 있듯이 데자뷰(DjVu)도 만능이 아니고 한계가 있다. PDF가 문서 중심이라면 데자뷰(DjVu)는 이미지 중심으로 특히 컬러 원문 이미지 서비스에 기술이 우수하다. 전자도서관 컨텐츠 구축 시 다양한 자료 원문 파일 포맷을 검토할 때 서울대학교에서 적용한 사례를 참고한다면 시행착오를 줄이는 방법이 될 수 있을 것이다.

2. 데자뷰 포맷의 도입 배경

서울대학교중앙도서관은 1999년부터 서울대 학위논문 4만여 건의 원문을 구축하여 원문을 서비스 하여왔으며 2001년부터 3년간 다량의 컨텐츠를 구축하여 2단계 전자도서관 사업을 추진하고 있다. 컨텐츠 구축은 도서관 자료와 캠퍼스內 전 기관에 컨텐츠 구축 수요 조사를 실시하여 컨텐츠 구축 선정 위원회에서 구축 대상을 선정하였다. 컨텐츠 구축 대상 자료는 책자형 자료, 슬라이드, 마이크로필름, 사진, VOD, AOD 등으로 자료의 유형이 다양

하다. 전자도서관 소프트웨어의 개발과 고문헌, 슬라이드 및 사진자료를 고해상도 컬러이미지로 대량의 컨텐츠 구축 사업을 추진하다 보니 원문의 파일 크기가 커서 서비스에 대한 다음과 같은 사업추진의 문제점이 발생하였다.

○ PDF 파일의 한계

책자형 이미지는 보통 PDF 포맷으로 원문을 구축하는데 파일크기가 10MB 이상이면 PDF파일이 열리지 않으며 10MB 이하라도 파일이 큰 PDF 파일은 원문을 보는데 시간이 많이 소요된다.

○ 네트웍 트래픽

파일 사이즈가 큰, PDF 파일, 슬라이드 및 사진 등의 용량이 큰 컬러 이미지는 이용자가 많을 경우 네트웍의 트래픽을 유발할 수 있다.

○ 고해상도 유지를 위한 적정 해상도 유지

원문 이미지의 고해상도 유지를 위하여 컬러로 구축되는 이미지의 서비스에 장애가 되지 않는 적정한 해상도 유지가 어렵다.

○ 저장장치 확장비용 및 Back-up

텍스트 데이터와 달리 이미지 파일은 사이즈가 커서 컨텐츠구축에서 저장장비 확장 및 Back-up 비용을 최소화 할 수 있어야 하고 용이하여야 한다.

위와 같은 문제점을 해결하기 위하여 서울대학교에서는 데자뷰를 분석하여 제한적으로 도입하게 되었다.

3. 데자뷰 포맷

가. 데자뷰란?

DjVu라는 단어는 불어 'déjà vu' 에서 온 것으로 '어디서 본 듯한 느낌 또는 착각' 이란 뜻을 갖고 있다고 한다. 우리말로는 '데자뷰'로 발음한다. DjVu는 웹상에서 통용되는 TIFF, PDF, JPEG과 같이 또 다른 하나의 파일 포맷이기도 하다. DjVu로 만들어진 문서는 확장자가 'djvu' 또는 'djv'로 표시된다. DjVu는 원문 이미지 구축시 원본자료의 그 품질 그대로 유지하고 검색, 배포, 압축, 저장할 수 있게 한 압축 기술이다. 이 기술은 1990년대 후반에 미국 AT&T Lab에서 연구 개발된 것으로 2000년 초에 미국 시애틀 소재의 리자드텍사에서 관련 기술을 인수하면서 보급되기 시작한 솔루션이다.

DjVu 기술이 주는 주요 사상은 "Scan-to-Web' 이다. 그 동안 스캐닝을 통해서 디지털화되는 일반 문서나 고화질의 사진 등이 그 파일의 크기가 커서 웹을 통해서 서비스가 불가능하던 문제점을 해소하는 기술인 것이다. 최근에는 고품질의 컬러 스캐너와 디지털 카메라의 보급으로 인하여 오프라인상의 자료들에 대한 디지털화 요구가 증대되고 있는 시점에서 이러한 기술은 진가를 발휘하게 된다.

DjVu로 표시된 문서도 PDF문서와 마찬가지로 자유롭게 웹에서 저장, 배포, 다운로드, 이메일 전송 등이 가능하다. 1,000페이지가 넘는 컬러 책자를 300 DPI 해상도로 스캐닝 하여 웹에서 서비스한다고

할 때 300 DPI 해상도의 품질을 웹에서 유지하기 위해서는 적어도 400 Mb (JPEG으로 페이지당 500Kbyte 정도 소요)이상의 파일 크기가 필요하게 된다. 400 Mb 나 되는 책자를 웹에서 서비스하는 것은 거의 불가능하다고 볼 수 있을 것이다.

DjVu는 400 Mb를 1/20로 압축하여 20 Mb 만들 수 있고 페이지 단위의 스트리밍 기술을 적용하여 책자의 페이지수와 파일의 크기에 상관없이 고화질의 원문을 일정한 검색 속도(1-2초 정도)를 유지하면서 이용할 수 있게 한다.

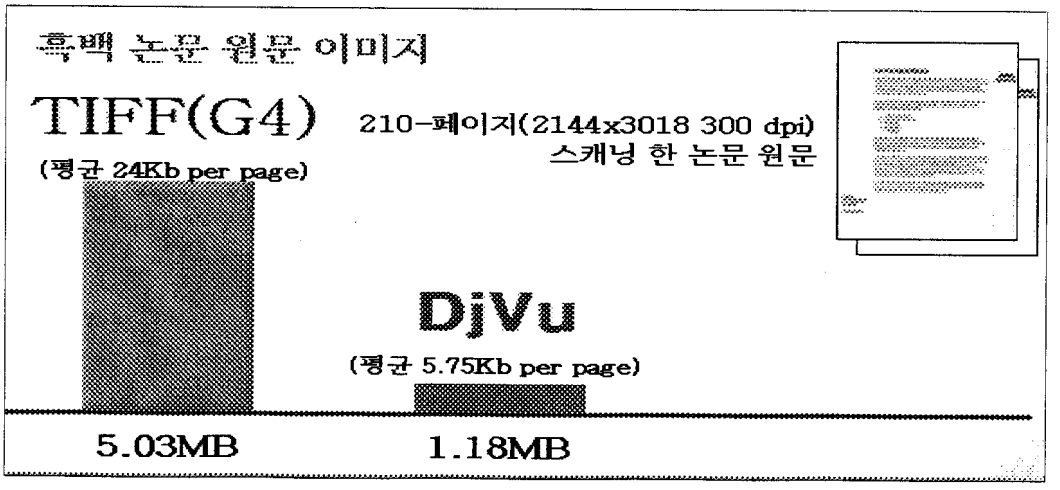
지도나 도면을 디지털화 할 경우 한 장의 파일 크기는 100 Mb 또는 200 Mb 이상 넘는 경우가 보통이다. 이 정도 크기면 일반적인 방법으로 웹에서 서비스할 수 있는 정도의 크기를 넘어서게 된다. DjVu는 이러한 지도나 도면과 같이 한 장의 이미지의 크기가 큰 경우도 적절한 크기로 압축을 할 뿐만 아니라 웹에서 서비스할 경우 소위 점진적인 이미지 전송 기술을 통해서 아무리 큰 크기의 이미지라도 사용자의 화면에는 기다림 없이 이미지를 순식간에 볼 수가 있게 하여 준다. 또한 스캐닝 또는 디지털 카메라를 통해서 만들어진 모든 디지털 자료나 사진들은 DjVu로 변환이 가능하다.

나. 원문 이미지 포맷과 데자뷰 포맷 비교

(1) TIFF 포맷과 DjVu와의 비교

1980년에 CCITT 표준화 그룹에서는 소위 Bi-level 이미지(흑백 이미지)를 표현하기 위한 Group 3 표준 포맷을 정했다. 1984년에는 G3 표준을 좀더 향상시킨 압축 권고안 Group 4가 발표되었다. 팩스 전송뿐만 아니라 디지털 도서관 분야에서 흑백 원문을 스캐닝 하여 이미지화하는데 가장 많이 사용하고 있는 포맷이 바로 G4 표준 압축 방식으로 표현되는 TIFF G4 타입이다.

G4 표준이 발표된 이후 1993년에 Joint



[그림-1]

Bi-level Images Experts Group (JBIG)에서는 새로운 흑백 이미지 코딩 표준을 JBIG1 이름으로 발표하였으나 G4 표준보다 좋은 압축률을 갖고 있음에도 불구하고 G4 표준만큼 널리 보급되지는 못하였다. 2000년에는 G4 표준 보다 약 3-4배 정도의 압축률이 좋은 JBIG2가 발표되었다. AT&T 는 당시에 JBIG2 표준에 근거하여 흑백 문서이미지 압축 포맷을 개발하였는데 이것이 DjVu JB2이다.

DjVu JB2 포맷은 일반적으로 TIFF G4 이미지 크기보다 약 3-5배 정도의 압축률 갖고 있다. [그림-1] 에서 보는 거와 같이 300 DPI 해상도로 스캐닝 한 210 페이지 학위 논문 이미지를 TIFF G4 크기와 DjVu 포맷의 크기를 비교 할 수 있다. 대학에서는 보통 학위논문의 해상도를 150 DPI로 서비스하고 있어 원문이 다소 선명하지가 않다.

TIFF G4는 단순히 흑백 이미지를 압축하는 표준 포맷일 뿐이다. TIFF 포맷 안에 PDF 문서와 같이 목차를 보여주는 책갈피 정보를 삽입하여 활용할 수 있게 하는 기능, 메타 데이터를 이미지 안에 삽입하여 검색 시 활용하는 기능, 원문의 본문 내용을 검색할 수 있도록 텍스트 정보를 원문이미지와 같이 표현할 수 있는 기능 등을 지원하지 않기 때문에 이러한 기능을 위해서는 별도의 응용프로그램을 만들어서 제공해야 하는 문제점들을 갖고 있다.

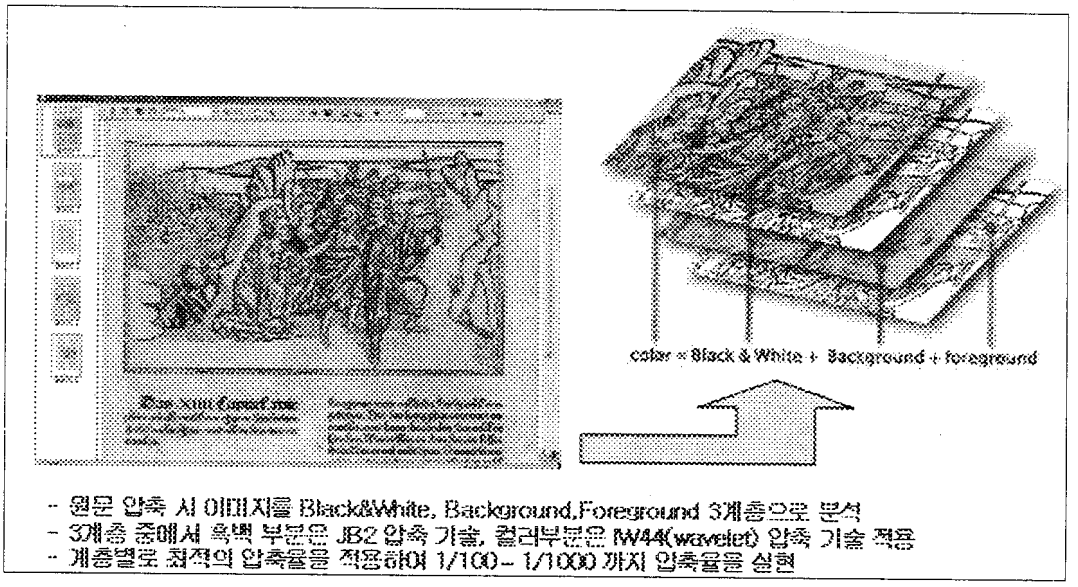
다른 이미지 포맷과는 다르게 여러 장의 페이지를 하나의 파일에 번들로 저장할 수 있는 Multi-TIFF 기능을 제공하기도 하지만 위에서 언급한 기능들을 위한 응용프로그램에서는 이러한 Multi-TIFF 기능과 함께 활용할 수 없기 때문에 별로 효용성이 없는 것이 되어 버렸다.

TIFF는 표준 포맷으로 인식되어 있음에도 불구하고 익스플로러와 같은 표준 브라우저에서 직접 TIFF 문서를 볼 수 없기 때문에 별도의 TIFF viewer를 설치하여 이용하여야 하는데 TIFF viewer 기능도 표준화되어 있지 않아 서비스하는 기관이나 업체에 따라서 그 사용법과 서로 호환이 안 되는 문제점을 갖고 있기도 하다.

DjVu는 흑백 이미지와 컬러 이미지에 대해서 각각 특성에 맞게 서로 다른 압축 방식을 사용하는 것이 특징 중에 하나이다. 흑백과 컬러가 혼합된 복합 이미지 문서는 DjVu로 변환 시 이미지 처리 기술을 이용하여 흑백 부분과 컬러 부분을 자동적으로 분리하여 각각 다른 압축 방식을 적용하기도 한다.[그림-2 참조]

DjVu의 흑백 압축 방식은 앞에서도 언급한바 와 같이 CCITT G4 방식 보다 4-5배 정도의 압축률을 제공하는 ISO 표준 규격 JBIG2에 준하여 개발된 JB2 방식을 사용한다. JB2 방식의 가장 특징은 흑백 이미지 문서들을 압축 할 시에 문서 페이지 전체 이미지를 대상으로 압축하는 방식을 사용하지 않는다.

이미지 처리 기술을 이용하여 중복해서 반복되는 문자 패턴을 자동으로 추출하고 이를 같은 그룹으로 분류하는 방식을 사용한다. 분류된 그룹은 대표가 되는 이미지만 압축 저장이 되고 나머지 대상은 위치 정보만을 갖게 한다. JB2는 한 페이지 내에서 추출한 압축 정보를 다음 페이지



[그림-2]

를 압축할 때 정보가 공유할 수 있도록 하였기 때문에 압축대상이 되는 페이지 수가 많을수록 압축률의 진가를 발휘하게 된다.

(2) JPEG 포맷과 DjVu와의 비교

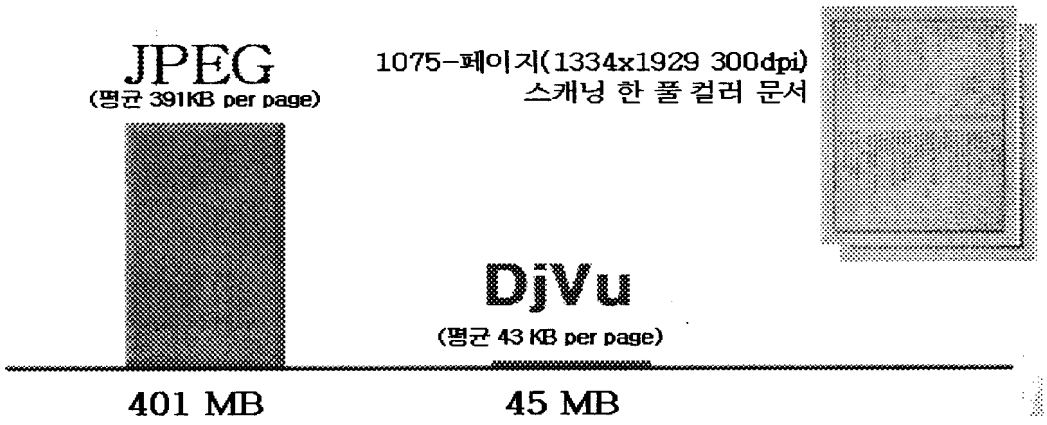
컬러 이미지를 저장하는 방식 중에 가장 보편화된 포맷이 JPEG이다. JPEG은 나름대로 압축 방식을 제공하기 때문에 컬러 이미지 파일을 표현하고 저장하는데 매우 적당한 포맷으로 인식되고 있다.

JPEG 보다 압축률을 향상시킨 JPEG2000 표준안이 2000년에 발표가 되었다. JPEG 압축방식과 다른 Wavelet 알고리즘을 기반으로 하여 만들어진 표준안이다. 효율성이나 성능, 이미지 품질 면에서 JPEG 보다 훨씬 좋은 것으로 평가되고는 있으나 아직까지 JPEG2000이 보편하게 이용되지 않고 있는 실정이다.

JPEG은 일반적으로 한 장으로 표현되는 사진과 같은 데이터를 표현하고 저장하는데 적당하다고 할 수 있으나, 컬러로 자료를 스캐닝 하여 여러 장을 하나의 책으로 묶어서 저장하고 서비스하는 구조에는 잘 맞지 않은 단점이 있다. TIFF 포맷과 같이 여러 장의 이미지를 한 개의 번들로 묶을 수 있는 Multi-TIFF와 같은 방식이 지원되지 않기 때문이다.

DjVu는 컬러 이미지 압축을 위해서 JPEG2000과 같은 Wavelet 기반의 압축 방식인 IW44 방식을 이용한다. 압축률은 JPEG과 비교하여 같은 질을 보장하면서 그 크기는 JPEG보다 1/10 - 1/20 정도로 압축이 될 수 있다. [그림-3] 참조.

DjVu는 JPEG 포맷과 마찬가지로 낱장 단위로 압축 저장할 수 있지만 Multi-TIFF방식과 같이 여러 장의 이미지 문서를 하나의 파일로 묶어서 저장 및 서비스를 할 수 있다. 원본의 이미지가 고해상도의



[그림-3]

이미지 인 경우 DjVu로 압축된 이미지라고 할지라도 파일 크기가 클 수 가 있다. 그러나 웹에서 일반 브라우저에 플러그인 된 DjVu 전용 viewer를 통해서 보게 될 경우 TIFF나 JPEG과 같이 전체의 파일을 모두 다운로드 받은 후 디스플레이 하는 방식을 이용하지 않고 점진적인 이미지 전송 기술을 이용하기 때문에 사용자는 원하는 이미지를 지연 시간 없이 즉시 볼 수가 있게 된다.

[그림-4]에서 고 앨범 샘플을 보면 총 50페이지 되는 고화질(600 DPI)의 원본을 JPEG으로 표현하면 총 23.3 Mb가 소요되나 DjVu로 압축을 하면 1.85Mb가 된다. 또한 고문헌 샘플을 보면 총 918 페이지의 JPEG으로 표현하면 378 Mb가 되나 DjVu로 압축을 하면 181 Mb가 된다. 압축된 파일의 크기도 페이지 수가 918 페이

연속귀중본 (컬러 300DPI)

(KB)

	JPG	DjVu
00000001	224	130
00000002	180	102
00000003	228	164
00000004	192	125
00000005	212	141
00000006	203	131
00000007	227	144
00000008	181	113
00000009	240	168
00000010	213	138
total	2,040	1,350

서울대 고문헌 샘플

총페이지수	압축형식	Total Size	페이지 당 평균 크기
918	JPEG	378.20 MBytes	432 KB
	DjVu	181.57 Mbytes	207 KB

고 앨범

총페이지수	압축형식	Total Size	페이지당 평균크기	특 징
50	JPEG	231.3 Mbytes	4,852 KB	600 DPI
	DjVu	1.85 Mbytes	39 KB	Scan option
	DjVu	11.8 Mbytes	248 KB	Photo option

[그림-4]

지나 되기 때문에 180Mb를 넘게 된다. 그러나 웹에서 180Mb나 되는 DjVu 문서를 검색하여 보면 전체 180 Mb를 모두 다운로드 받아서 디스플레이 하지 않고 필요한 페이지만 선택적으로 스트리밍을 해서 보여 지기 때문에 지연 시간 없이 볼 수가 있게 된다.

(3) PDF 포맷과 DjVu와의 비교

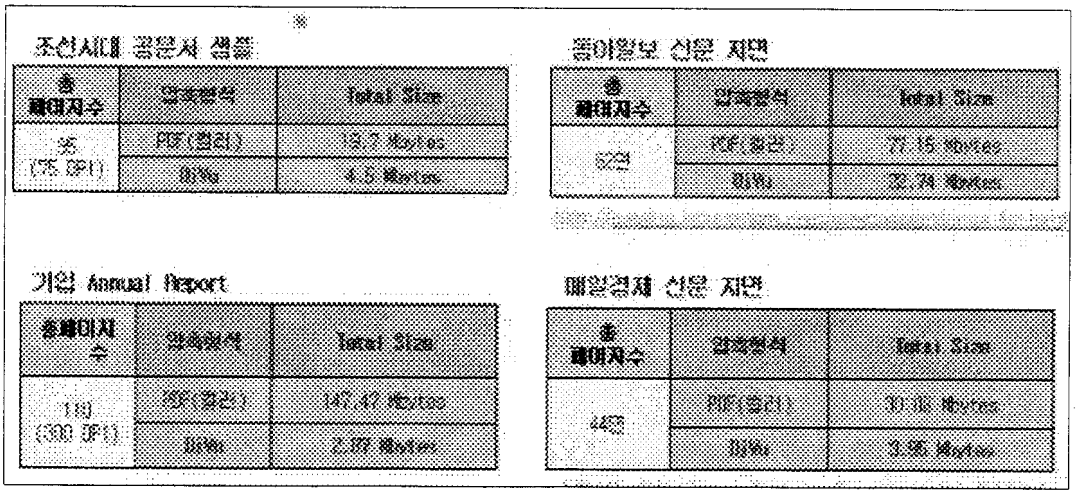
PDF 포맷은 본래 디지털 문서들을 웹을 통해서 Publishing 하기 위한 솔루션으로 이용되어 왔다. 워드나 아래아한글 등으로 작성된 텍스트 기반의 문서를 웹을 통해서 배포하기 위해서는 PDF 포맷만큼 좋은 툴은 없을 것이다. PDF 는 전용 Acrobat Reader를 무료로 사용하면서 누구나 쉽게 PDF 문서를 읽을 수 있는 장점이 있다. 뿐만 아니라 PDF 문서는 내부 문서를 수정할 수 없고 읽을 수 만 있게 되어 있기 때문에 더더욱 문서 배포용으로는 매우 적당한 솔루션이다. 그러나 스캐닝 한 이미지 문서를 PDF 포맷으로 저장하고 배포하는 데는 그 한계점이 180 Mb 정도이다.

스캐닝 한 문서는 그 자체가 이미지 포맷으로 디지털화 된 텍스트 문서보다는 파일 크기가 몇 배나 클 수밖에 없다. PDF는 기본적으로 압축을 기반으로 하지 않고 있기 때문에 스캐닝 된 이미지 문서를 PDF로 담기 위해서는 그 크기 이상의 파일을 요구하게 된다.

○ 컬러 문서의 PDF 포맷과 DjVu

기업보고서 110페이지 책자를 300 DPI 컬러로 스캐닝 하여 PDF로 저장을 하게 되면 약 140Mb 정도 소요가 되나. DjVu로 저장을 하면 약 3Mb 정도로 가능하다.

인터넷에서 신문 지면 서비스용으로 PDF 포맷이 많이 이용되고 있다. 그러나 [그림-5]에서 알 수 있듯이 신문 66면 정도의 지면에 컬러 광고까지 모두 포함하면 PDF 문서의 크기는 무려 70Mb가 넘게 된다. 따라서 각 신문사는 신문 지면



[그림-5]

A comparison by James Rile, PlanetDjVu, Sept. 19, 2002, updated Oct. 22, 2002

Document	PDF Image with Group 4 Compression	Searchable Image with Group 4 Compression	DjVu Image using JB2 Compression	Searchable Image DjVu using JB2 Compression
Contract	56k	72k	**10k**	**12k**
Annual Report	2.89M	2.0M	**396k**	**527k**
Technical Report	2.02M	2.15M	**368k**	**419k**
Patent	815k	906k	**123k**	**162k**
Total	5.84M	5.13M	**897k**	**1.12M**
Average	100%	100%	**15%**	**22%**

참조 : http://www.planetdjvu.com/comparison_of_advanced_compression_for_bitonal_images.htm

[그림-6]

서비스를 위해서 66면을 하나의 PDF파일로 묶어서 서비스하지 못하고 각 면 단위로 PDF파일로 서비스한다.

일반 디지털 문서를 위한 포맷으로 PDF가 적당하다면 DjVu는 컬러로 스캐닝한 이미지에 대해서는 PDF 포맷 보다 훨씬 뛰어남을 알 수 가 있다.

○ 흑백 문서의 PDF 포맷과 DjVu

스캐닝한 흑백문서를 PDF 포맷과 DjVu와 비교하여 보면 DjVu 포맷이 PDF 문서보다 평균 1/4로 압축된다. 압축비율이 스캐닝한 해상도에 따라 차이가 나며 낱장문서나 페이지가 적은 자료는 PDF 포맷이 좋지만 몇 백 페이지가 넘는 자료는 데자뷰 포맷이 효과가 있다.[그림-6]

다. 데자뷰의 장단점

앞에서도 데자뷰 포맷에 대하여 언급하였지만 데자뷰 포맷의 장단점을 요약하면 아래와 같다.

(1) 장점

○ 일반적으로 많이 사용하는 PDF는 문서전용 포맷이지만 DjVu는 이미지 전용 포맷이다.

○ 고해상도 컬러 이미지 원문의 경우 고품질 해상도를 유지하면서 압축률이 높아 파일 사이즈를 최소화 할 수 있고 원문 저장 장치의 확장 비용을 절감할 수 있다.

○ 페이지 스트리밍이 지원되어 큰 용량의 파일이라도 필요한 페이지만 가져와서 보여 주기 때문에 1-3초 이내 원문을 볼 수 있어 원문 서비스의 속도를 개선할

수 있고 네트웍 트래픽을 줄일 수 있다.

○ 사이즈가 큰 파일은 PDF 뷰어에서 열리는데 한계가 있지만 DjVu는 파일 사이즈에 관계없이 서비스 할 수 있다.

○ 원문보기에서 원문을 확대하여 보아도 이미지의 손실이 없이 확대하여 볼 수 있다.

(2) 단점

○ 별도의 DjVu 전용뷰어가 필요하다. PC에 데자뷰 전용 뷰어가 설치되어 있지 않았을 때 몇 십초 이내에 자동으로 설치되지만 원문 컨텐츠에서 일반적으로 많이 사용하는 PDF 뷰어 이외의 또 하나의 뷰어를 사용하게 된다.

○ PDF 뷰어보다 DjVu 뷰어가 원문보기에 있어 다소 불편하다.

○ 데자뷰 포맷을 이용하기 위해서는 최소 100,000면 단위로 라이센스 비용을 지불하여야 한다.

○ 스캔 된 문서가 아닌 기존 워드프로세서 문서 (MS-word, 흔글, 엑셀 등)로 작성된 문서를 Djvu로 변환 시 변환 과정이 번거롭고, 변환 후에도 압축 효과를 기대할 수 없다.

○ 데자뷰 파일은 PDF와 같이 원문 파일을 사용자가 수정 편집할 수 없다. 책자자료는 문제가 없지만 사진이나 슬라이드의 낱장자료는 필요에 따라 편집을 하여 사용할 수 있어야 하나 편집이 불가능하다. 이는 낱장 사진자료 적용에 있어 단점이 되기도 한다. 그러나 저작권의 보호측면에서는 장점이 될 수도 있다.

4. 데자뷰 포맷 서울대학교 적용 사례

가. 원문이미지 구축의 기본방침

서울대학교는 원문 컨텐츠 구축에 있어 다음과 같은 기본 방침 아래 구축을 하고 있다.

○ 중요자료의 원본 보존

고문헌 같은 중요자료는 보통 마이크로필름으로 촬영하여 보존하지만 한번 스캔한 자료를 원본상태와 거의 유사한 해상도로 스캔하여 원본 보존을 이미지로 대체한다.

○ 고품질 원문 구축

한번 구축한 원문의 재구축을 피하기 위하여 원문구축은 현재의 장비나 기술 중 가장 우수한 품질로 구축한다.

○ 원본 포맷과 서비스 포맷 동시 보존

이미지, AOD, VOD 등 구축되는 모든 컨텐츠는 고품질로 일차 원본을 구축하고, 서비스는 압축한 포맷으로 서비스한다. 따라서 원본 포맷과 서비스용 포맷이 달라지기도 한다.

○ 서비스 포맷의 유동적 대처

원본 파일을 고품질로 구축하고 보존하여 서비스 포맷의 기술이 발전되면 저장된 원본 파일로 새로운 서비스 포맷으로 변환하여 서비스한다. 예를 들면 컬러 이미지 원문의 경우 현재는 PDF 보다 데자

뷰 포맷이 우수하여 데자뷰 포맷으로 서비스 하지만 PDF 포맷이 데자뷰보다 더 우수하거나 새로운 우수한 포맷으로 업그레이드 될 경우 저장된 원본을 가지고 PDF 포맷이나 새로 개발된 우수한 포맷으로 변환하여 서비스한다.

나. 데자뷰 포맷의 적용대상

원문 컨텐츠 구축에 있어 다음의 경우는 서비스 포맷에 한하여 데자뷰로 변환하여 서비스한다.

(1) 컬러이미지 원문구축자료

컬러 이미지로 구축하는 원문의 원본은 300 DPI JPEG으로 보관하고, 아래 자료의 서비스 포맷은 데자뷰 포맷으로 변환된 각 DPI(300, 200, 150)별로 파일 사이즈와 해상도 등을 다각도로 검토했을 때, 가장 적정선이 보장되는 200 DPI로 데자뷰 포맷으로 서비스한다.

- ㅇ 고문헌
- ㅇ 조선 근대 신문
- ㅇ 창간호 잡지
- ㅇ 슬라이드 자료
- ㅇ 사진 및 필름자료
- ㅇ 미술 작품집

(2) 흑백 이미지 원문구축자료

흑백 이미지로 구축한 컨텐츠 중 파일 사이즈가 큰 다음의 컨텐츠는 데자뷰 포맷으로 서비스한다.

- ㅇ 탁본
- ㅇ 대학사료
- ㅇ PDF 파일로 구축한 자료 중 사이즈가 큰 자료

다. 데자뷰 포맷 적용방법

(1) 원본 보정작업 후 일괄 변환

세계적으로 데자뷰 포맷을 적용하여 원문 컨텐츠를 적용하고 있는 기관이 있지만 국내에서 처음으로 서비스 포맷에 데자뷰를 도입함으로써 시행착오를 줄이기 위해 여러 방면으로 연구, 검토하였다. 컬러 이미지를 300 DPI TIFF로 스캔하면 사이즈가 커서 이미지 보정 작업에 어려움이 대두되었다.

따라서 스캔은 300 DPI JPEG으로 하여 1차로 이미지 보정이 필요한 것은 이미지를 보정한다. 이미지 보정이 완료되면 JPEG 자료의 종류별로 일괄 데자뷰로 변환한다. 데자뷰로 변환하는 방법은 구축 시 하나하나 변환하는 방법이 있고, 대량의 데이터를 일괄로 변환하는 방법이 있지만 서울대에서는 고문헌, 슬라이드, 미술작품집 등 매체 구축별로 일괄 변환하였다.

(2) 메타데이터와 연계 및 TOC 북마크 생성

데자뷰 변환 시 결여되었던 TOC 북마크 기능은 보완하여 생성하고 이용을 위해 데자뷰로 변환 된 파일들과 별도로 구축 된 각종 메타데이터(서지, 관리, 저작권)와 연계는 구축자료 업로드 시 서로 제어번호로 연동 관계를 맺고 서비스한다.

(3) 데자뷰 적용 해상도

- ㅇ 고문헌

페이지당 300 DPI JPEG으로 스캔하여 200 DPI DjVu로 서비스한다.

○ 슬라이드

페이지당 2400 DPI JPEG으로 스캔하여 600 DPI DjVu로 서비스한다.

(4) 데자뷰 서비스 화면

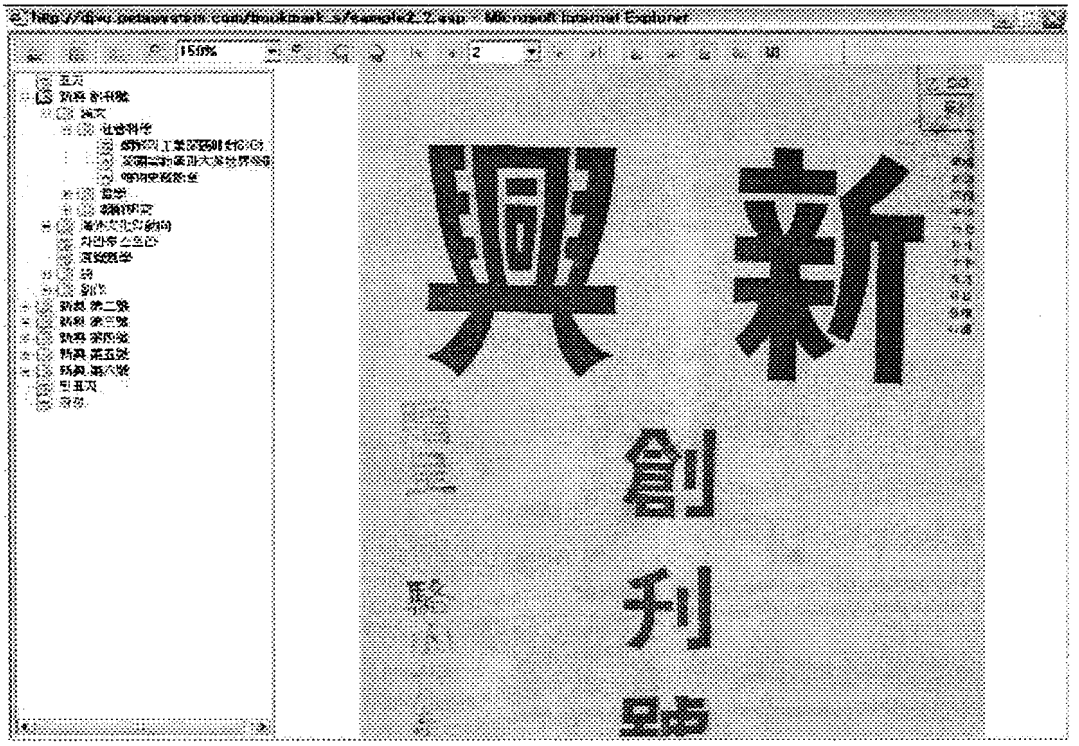
[그림-7]은 서울대학교 중앙도서관의 고문헌 중 웹에서 익스플로러 브라우저에 플러그인 된 DjVu 전용 viewer를 통해서 검색한 결과 샘플 화면이다. 상단에는 메뉴 툴바가 있고 좌측에는 목차 정보를 책갈피 형태로 나타나게 하였다. 축소 확대를 자유롭게 할 수 있어서 원하는 부분을 자세히 볼 수도 있다.

라. 데자뷰 포맷 적용시 제기되었던 문제점

국내에서 처음으로 원문 서비스 포맷으로 데자뷰를 적용하면서 여러 가지 문제점이 제기 되기도 하였다. TOC의 파일의 북마크가 생성되지 않는 다든가 원하는 페이지만 프린트하는 기능 등 여러 가지 문제점이 대두되었지만 이를 개발하여 문제점을 해결하였으나 저작권 보호를 위한 DRM은 아직 적용되지 않는다.

마. 기대효과

컬러 이미지 원문과 흑백 이미지 원문



[그림-7]

의 일부를 데자뷰로 적용하여 고품질 해상도로 원문을 서비스 할 수 있다. 원문보기에서 파일 사이즈에 관계없이 원문을 빠른 속도로 볼 수 있고 다수의 이용자가 원문을 이용하더라도 네트웍의 트래픽을 최소화 할 수 있을 것으로 기대된다. 또한 대량의 원문 컨텐츠를 구축하면서 원문 저장장비를 1/3로 축소 할 수 있었다.

5. 결 론

국내 도서관에서는 1990년대 중반부터 일부기관에서 원문이미지를 구축하기 시작하였다. 1990년대에 구축한 원문 이미지 포맷은 주로 TIFF 이었다. TIFF는 표준 포맷으로 인식되어 있음에도 불구하고 익스플로러와 같은 표준 브라우저에서 직접 TIFF 문서를 볼 수 없기 때문에 별도의 TIFF viewer를 설치하여 이용하여야 하는데 TIFF viewer 기능도 표준화되어 있지 않기 때문에 서비스하는 기관이나 업체에 따라서 그 사용법과 서로 호환이 안되는 문제점을 갖고 있었다.

그 후 PDF가 보편화되면서 PDF 포맷으로 원문 컨텐츠를 구축하여 서비스하고 있지만 문서 중심의 PDF 포맷으로는 고품질로 다양한 유형의 원문 컨텐츠를 서비스하기에는 현재로서는 한계가 있다. 서울대학교에서는 이러한 한계를 극복하기 위하여 일부 컨텐츠는 서비스 포맷을 DjVu로 적용하였다. 문서중심의 PDF와 이미지 중심의 DjVu를 적절히 원문 서비스 포맷에 적용하지만 그 효과는 좀더 사용하여 보아야 할 것이다.

금년 여름에 발표 예정인 Acrobat Reader6.0에서도 DjVu와 같은 페이지 스트리밍 기술이 적용되어 큰 사이즈의 원문파일이 PDF에서 열리지 않는 것은 해결될 수 있을 듯하지만 컬러 원문 이미지의 고품질 서비스는 당분간 DjVu가 우수하여 이 포맷을 적용할 수밖에 없을 듯하다. 다만 데자뷰 포맷을 적용시 원본 파일과 서비스 파일의 포맷을 따로 하여 원본 파일을 보관하는 것이 기술발전에 적응하는 방법일 것이다.

데자뷰는 원본 파일의 포맷으로 보다 서비스 포맷 파일로 사용하여야 또 다른 우수한 포맷이 나온다면 원본 파일로 새로운 다른 포맷으로 변환하여 원문을 서비스 할 수 있을 것이다.