다양한 문서형식을 지원하는 기술정보은행(DFR) 개발

†손원성O, ††남동선, †††임순범

†경인교육대학교 컴퓨터교육과
††(주)한글과 컴퓨터
†††숙명여자대학교 멀티미디어학과

†sohnws@ginue.ac.kr, ††speeno,@haansoft.com, †††sblim@sookmyung.ac.kr

Implementation of Digital Format Registry for Opened Office Document format

†Won-Sung Sohn,†† Dong-Sun Nam, †††Soon-Bum Lim

†Dept. of Computer Education, Gyeonin National University of Education
††Hansoft
†††Dept. of Multimedia Science, Sookmyung Woman's Univ.

요 약

본 연구에서는 XML 기반의 Digital Format Registry를 고도화하며 국내 실정(750 여개의 기록관)에 맞는 기술정보은행 서비스 제공 모델을 연구하여 적용방안을 수립하고 이에 필요한 요소 기술을 개발하는 것이다. 특히 도 고도화 연구 개발 사업에서는 이러한 표준에 대한 지원을 최우선으로 진행되었다. 특히 업무 환경에 가장 빈번히 사용되는 국제 표준 기반의 ODF(Open Document Format)와 OOXML(Office Open XML)이라는 개방형 오피스 문서 형식을 추가 지원하였으며, ISO 28500 WARC라는 웹 기록물 저장 국제 표준 형식에 대한 유효성 검증 기능을 개발 추가하여 웹 기록물에 대한 장기보존 시스템의 신뢰성을 높일 수 있도록 하였다.

1. 서론

"디지털화 된다."는 말은 기존의 종이 방식으로 되어 있는 문서를 어떤 형식의 바이트 스트림으로 디지털 객체로 변환하는 것을 말한다. 이러한 종이 방식의 문서를 디지털 객체로 변환할 때, 이 객체의 특정한 형식을 디지털 포맷이라고 한다.

일반적으로 컴퓨터를 활용하여 생성된 전자데이터의 형식은 앞에서 말한 특정한 디지털 포맷 규칙을 통해 디지털 객체로 저장되고, 그 디지털 객체를 생성하고, 수정하고 사용자가 인식할 수 있도록 구체화하는 방법은 특정한 소프트웨어를 활용하거나 하드웨어 환경 하에서 수행되는 것이 일반적이다. 이러한 디지털 정보는 공공기관, 교육기관, 대용량 정보처리기 등에서는 장기적으로 보관되어야 하며 오랜 세월이 흘러 해당 디지털 객체를 활용함에 있어서 원본에 대한 재현이 가능하여야 한다.

따라서 디지털 보존, 아카이브, 아카이빙 관련 시스템 및 이를 구축하기 위한 디지털 정보의 보존에 관한 정책, 표준화, 방법론, 보존기능, 절차 등에 대한 다양한 연구가 진행되고 있다[1,2,3,4]. 특히 디지털 문헌의 보존 문제를

※ 본 연구는 행정안전부 국가기록원의 지원을 받아 기록물 보존기술 연구개발(R&D) 사업의 일환으로 이루어졌으며, 이에 감사드린다.

해결하기 위한 포맷 레지스트리(Digital Format Registry)에 관한 연구는 다양한 형태로 진행되고 있다[,5,6,7,8]

포맷 레지스트리란 일반적으로, 특정 디지털 정보 파일의 포맷 구문정보와 의미정보를 저장하는 일종의 데이터베이스이며, 특정 어플리케이션이나 기술적인 환경 변화가 일어나도 디지털 정보를 장기간 보존 할 수 있게 하는데 목적이 있다. 그 결과 DFR(Digital Format Registry)은 장기 보존 전략에 관한 주요한 기술이며 국가적 단위의 대규모 프로젝트 형태로 다양한 연구 및 개발이 진행되고 있다. 그러나 국내에서는 본 기술정보은행 시스템과 같이 장기 보존을 위한 DFR 시스템을 구현한 사례가 없을 뿐 더러, 연구 또한 제대로 이루어지지 않고 있다.

따라서 본 논문에서는 국내 환경에 적합한 “디지털 포맷 기술 정보 은행 시스템”을 설계하고 구축하도록 한다. 또한 국내의 디지털 객체 활용 여건(750여 개의 디지털 객체의 종류)과 레거시(Legacy) 디지털 객체의 처리 기술을 개발하여, 보다 국내 실정에 적합한 DFR 시스템을 구축하였다.

또한 본 연구에서는 이러한 표준에 대한 지원을 최우선으로 진행되었다. 특히 업무 환경에 가장 빈번히 사용되는 문서 형식으로 2006년, 2008년에 각각 ISO 26300과 ISO 29500 이라는 국제 표준 승인된 ODF(Open Document Format) - (추가 15종의 문서 형식)과 OOXML(Office Open XML) - (추가 3종의 문서 형식)이라는 개방형 오피스 문서 형식을 추가 지원하게 되었으며, ISO 28500 WARC라는 웹 기록물 저장 국제 표준 형식에 대한 유효성 검증 기능을 개발 추가 함으로 웹기록물에 대한 장기보존 시스템의 신뢰성을 높일 수 있는 요소 기술로 사용될 수 있게 되었다.

또한 국내 실정을 반영하여 국가 기록원의 장기 보존 저장 형식인 NEO Package 파일에 대한 유효성 검증을 수행할 수 있게 됨으로 기존에 수작업으로 진행하여 업무의 효율과 정확도가 낮았던 보존 문서의 확인 절차를 자동화 및 간편화 할 수 있게 되어 비용과 시간을 줄일 수 있도록 하였다.

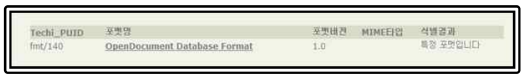
2. Open Document Format 지원을 위한 방안

오픈 도큐먼트 포맷(Open Document Format)는 스프레드시트, 차트, 프레젠테이션, 데이터베이스, 워드 프로세서를 비롯한 사무용 전자 문서를 위한 파일 형식이다. 이 형식은 원래 오픈오피스에서 만들고 구현한 XML 파일 형식을 바탕으로, OASIS(Organization for the Advancement of Structured Information Standards) 컨소시엄이 표준화하였다[9,10,12].

[그림 1] 기본 화면

기본적인 오픈도큐먼트 문서는 최상위 요소가 <office:document>인 XML 문서이며, 다른 파일들이 포함된 경우에는 ZIP 형식으로 하나의 파일로 압축할 수도 있다. 오픈도큐먼트는 내용과 스타일, 메타데이터, 프로그램 설정 등을 네 개의 XML 파일에 분리하여 저장함으로써 작업의 분리(SoC)를 구현한다.

이러한 ODF를 지원하는 DFR 구현화면은 [그림 1]과 같으며 식별과정에서의 결과 및 유효성 결과는 [그림 2], [그림 3]과 같다.



[그림 2] 식별 기능에 대한 결과 화면 확대

[그림 3] ODF 문서 식별 및 유효성 검사 결과 화면

3. Office Open XML 지원을 위한 DFR 개발방안

OpenXML[11]은 처음부터 Microsoft Corporation에서 정의한 이진 형식으로 인코딩된 기존의 워드 프로세서 문서, 프레젠테이션 및 스프레드시트 자료를 정확하게 표현할 수 있도록 하기 위해 설계되었다. 표준화 프로세스는 이러한 자료의 존속과 이들의 확장, 상세 문서의 제공 및 상호 운용성의 활성화를 나타내는데 필요한 기능을 XML에 반영하는 것으로 구성되어 있다.

OpenXML은 워드 프로세싱, 프레젠테이션 및 스프레드시트 문서의 형식을 정의한다. 각 문서 유형은 WordprocessingML, PresentationML 또는 SpreadsheetML과 같은 기본 표시 언어를 통해 지정된다. 포함 메커니즘을 통해 이 세 종류 중 한 가지 문서에 다른 기본 표시 언어로 된 내용과 지원하는 여러 표시 언어로 된 내용을 포함할 수 있다.

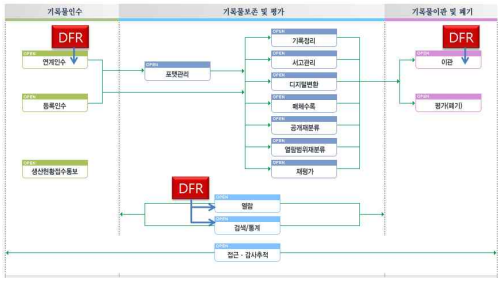
이러한 OOXML의 유효성 검사과정의 화면은 [그림 4] 및 [그림 5]와 같다.

[그림 4] OOXML 문서에 대한 식별 결과 화면

[그림 5] OOXML 문서 형식에 대한 식별 및 유효성 검증 결과 화면

4. 기술정보은행의 구현

본 연구에서 개발한 DFR의 구조는 [그림 6]과 같다.

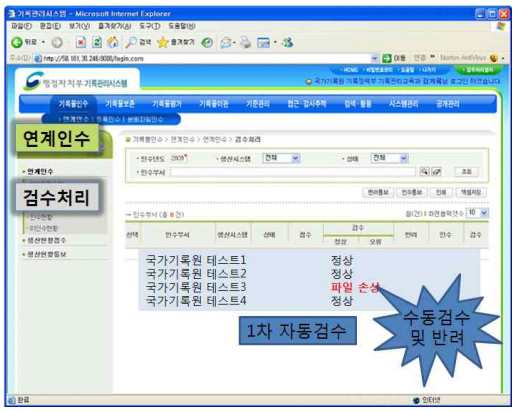


[그림 6] 본 과제 연구를 통해 도출된 DFR 시스템 적용 지점

본 연구를 통해 실제 국내 장기 보존 시스템 및 기록관리 시스템에서 얻을 수 있는 가장 큰 장점이 자동화 유효성 검증이라고 할 수 있다.

현재 기록물을 장기 보존 시스템에서 인수하는 시점에 이 전달된 디지털 객체에 대한 검수를 위해서는 모든 파일을 일일이 사람이 수동으로 열어봐야 하므로 현실적으로 검수에 인력과 시간이 많이 드는 상황이다.

이러한 비효율적인 업무 처리는 본 고도화된 DFR 시스템을 활용하여 1차적으로 자동검수로 자동적으로 검출되는 자료에 대해서 2차적으로 문제가 발생한 자료에 대해서만 사람이 재검수를 수행하여 업무처리의 효율성을 높일 수 있다. 이러한 과정은 [그림 7]과 같다.



[그림 7] 국내 기록관리 시스템 구현 방안 예

5. 결론

일반적인 사항으로 본 프로젝트의 대상인 DFR(Digital Format Registry)는 장기 보존 전략 분야를 연구하는 여러 국가에서 구현 및 관련 내용 연구가 진행되고 있으며, 한국에서 본 기술정보은행 시스템과 같이 DFR시스템을 구현한 사례가 없으며, 본 연구를 통해 연구 개발된 DFR시스템의 경우, 한국 내의 디지털 객체 활용 여건(처리 디지털 객체의 종류)과 레거시(Legacy) 디지털 객체의 처리 기술을 동시에 개발하였고, 국내외 적으로 DFR 시스템의 관점에서 포함되지 않았던 문서 디지털 객체의 텍스트 추출기능을 처음으로 구현하여 그 활용의 범위가 기존의 여타 DFR시스템보다 넓다고 할 수 있다.

6. 참고문헌

[1] Adrian Brown, "Automatic Format Identification Using PRONOM and DROID," http://droid.sourceforge.net/wiki/images/b/b4/Technical_Paper_1 Automatic_Format_ Identification_v2.pdf

[2] Stephen L. Abrams, "Establishing a Global Digital Format Registry," *Library Trends*, Vol. 54, No. 1, 2005.

[3] JHOVE, http://hul.harvard.edu/jhove/

[4] Metadata Extraction Tool Version 1.0 (National Library of New Zealand), http://www.natlib.govt.nz/en/whatsnew/4initiatives. html

[5] National Archives and Records Administration et al., Archival Workshop on Ingest, Identification, and Certification Standards, 2005.

[6] OASIS/ebXML Registry Information Model v2.5., http://www.oasis-open.org/committees /regrep/documents/2.5/specs/ebrim-2.5.pdf

[7] Searle, S., & Thompson, D., "Preservation metadata: Pragmatic first steps at the National Library of New Zealand," *D-Lib Magazine*, 9(4), 2003.

[8] Brown, Adrian, "Automating preservation: new develppments in the PRONOM service" *RLG DigiNews*, 9(2), 2005.

[9] ISO/IEC 26300:2006, Information technology -- Open Document Format for Office Applications (OpenDocument) v1.0, 2006

[10] Openoffice, http://www.openoffice.org, 2008

[11] Standard ECMA-376 Office Open XML File Formats, 2006

[12] StarOffice 8, http://www.sun.com/software/star/staroffice/index.jsp, 2008