Journal of Korea Multimedia Society Vol. 12, No. 10, October 2009(pp. 1397-1406)

디지털 객체 보존을 위한 디지털 포맷 레지스트리에 관한 연구

손원성[†], 임순범[††], 남동선[†††], 김은미[††††]

요 약

본 논문에서는 국내 환경에서 디지털 객체 보존에 관한 문제를 해결하기 위한 기술정보은행-디지털 포맷 레지스트리(Digital Format Registry)를 제안한다. 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.

본 연구를 통해 개발된 "기술정보은행" 시스템은 국내 환경에서 생성되는 다양한 전자기록의 장기보존과 장기 접근성 유지에 근간이 되는 기술정보요소(Technical Information)를 지속적으로 수집·관리하여 마이그레이션이나 에뮬레이션과 같은 보존 전략을 효율적으로 진행될 수 있도록 한다. 또한 제안 시스템은 소비자에게 필요한 정보를 쉽고 편리하게 검색할 수 있는 Key를 디지털 객체로부터 추출할 수 있도록 하는 기능 등을 제공하여 보다 효율적인 기록관리가 가능하도록 하였다.

A study on the Digital Format Registry for digital objects preservation in Korea

Won-Sung Sohn[†], Sun-Bum Lim[††], Dong-Sun Nam[†††], Eun-Mi Kim[††††]

ABSTRACT

This paper propose the "Digital Format Registry(DFR)" to solve the problem related digital objects preservation system in the Korean Industry. Digital format registry is a kind of database that saves syntax and meaning informations of a digital file format and for giving help to make preserve in long-term even technical environment of a specific application has been changing.

The role of the Technical Information Registry has been developed in this research and development is maintaining a technical information that is the foundation to maintain the long-term preservation and access to a digital objects. The function that can extract text information from a digital document object is implemented in DFR as a basic function at the first time in the world. This function make information consumers search a information that is needed easily and conveniently and can be used for development more effective records management system with retrieving the Key(index).

Key words: Digital Format Registry(디지털 포맷 레지스트리)

※ 교신저자(Corresponding Author) : 손원성, 주소 : 인천광역시 계양구 교대길 45(407-753), 전화 : 032)540-1284, FAX : 032)548-02881, E-mail : sohnws@gin.ac.kr
접수일 : 2009년 3월 18일, 수정일 : 2009년 6월 27일
완료일 : 2009년 7월 7일
[†] 종신회원, 경인교육대학교 컴퓨터교육과 조교수
[††] 종신회원, 숙명여자대학교 멀티미디어과학과 교수
(E-mail : sblim@sookmyung.ac.kr)
[†††] (주)한글과컴퓨터 연구개발실 선임연구원
(E-mail : speeno@haansoft.com)
[††††] 정회원, (주)한글과컴퓨터 연구개발실 XML응용기술팀 주임연구원
(E-mail : emkim@haansoft.com)
※ 본 연구는 행정안전부 국가기록원의 지원을 받아 기록물 보존기술 연구개발(R&D) 사업의 일환으로 이루어졌으며, 이에 감사드린다.

1. 서 론

정보기술은 개인이나 기관이 디지털 형식으로 문서나 지적생산물을 직접 생산하도록 하였으며, 기존의 많은 아날로그 문헌들을 쉽게 디지털로 재생산하는 도구를 제공하는 등 기록문헌의 디지털화에 많은 기여를 해왔다. 그 결과 현재 매우 많은 양의 디지털 문헌을 사회적 자원으로서 보유하게 되었지만, 이것을 아무런 문제없이 후대에까지 전승시킬 수 있는가 하는 의문에는 긍정적인 대답이 곤란한 상황이다. 그 이유는 다음 그림 1과 같이 종이와 같은 전통적인 매체보다 디지털매체의 문헌은 오랫동안 보존하는 것이 쉽지 않기 때문이다.

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. 특히 디지털 문헌의 보존 문제를 해결하기 위한 포맷 레지스트리(Digital Format Registry)에 관한 연구는 다양한 형태로 진행되고 있다[1-4]

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(처리되는 디지털 객체의 종류)과 레거시(Legacy) 디지털 객체의 처리 기술을 개발하여, 보다 국내 실정에 적합한 DFR 시스템을 구축하였다.

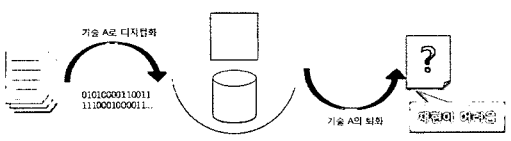


그림 1. 디지털 문헌 보존의 문제

본 연구에서 제안한 포맷 레지스트리의 역할은 전자기록의 장기보존과 장기 접근성 유지에 근간이 되는 기술정보요소를 지속적으로 수집·관리하여 마이그레이션이나 에뮬레이션과 같은 보존 전략을 효율적으로 수행할 수 있도록 지원한다.

그 결과 국내외 적으로 DFR 시스템의 관점에서 포함되지 않았던 문서 디지털 객체의 텍스트 추출기능 등을 제공하여 기존의 DFR 시스템보다 높은 활용성을 제공한다. 또한 기술 재활용에 염두를 두어 웹서비스 표준에 따라 그 기능을 구현 및 제공함으로써 디지털 기록물의 포맷을 처리 및 관리하기 위한 다양한 기능을 네트워크로 연결되어 있는 어느 곳에서든지 재활용할 수 있는 등의 장점을 제공할 수 있다.

2. 관련연구

정보기술과 인터넷의 비약적인 발전은 문헌자료를 인쇄매체가 아닌 디지털 자원으로 변화시키고 있다. 이러한 변화는 다양한 형태의 디지털 기록물에 대한 생성, 활용 및 보존을 필요로 하게 되었다. 일반적으로 컴퓨터에서 생성된 전자데이터는 특정한 디지털 포맷 규칙을 통해 디지털 객체로 저장되고, 그 디지털 객체를 생성하고, 수정하고 구체화하는 과정은 특정한 소프트웨어를 활용하거나 하드웨어 환경하에서 수행된다.

그 결과 전자 기록물은 특정 소프트웨어 및 하드웨어 기술에 의존적이며 기술의 변화 및 발전 속도가 빨라 관련 기술 및 디바이스가 퇴화될 경우 기존에 사용되었던 디지털 기록물은 재현될 수 없는 정보가 될 수 있다[5].

이러한 디지털 문서의 보존 문제가 대두됨에 따라 디지털 문서를 장기적으로 보존하고자 하는 다양한 노력이 시도되고 있다[6]. 국내외 각 기관들의 장기보존을 위한 절차들, 보존 방법론의 개발, 장기 보존의 표준화, OAIS 참고모형의 개발 등이 그 예이며[3], 그 중의 하나가 "디지털 포맷 및 애플리케이션 기술정보은행(DFR) 프로토타입 개발"에서 구현된 기술정보은행의 일종인 디지털 포맷 레지스트리이다 . 디지털 포맷 레지스트리는 포맷에 관한 일반적인 정보 뿐 아니라, 포맷에 관한 식별, 포맷의 유효성을 판단하는 검증, 포맷의 마이그레이션을 위한 중요

한 특징정보를 추출하는 특성정보 등의 기능을 제공하여 포맷의 장기보존에 중추적인 역할을 담당하고 있다[2]. 장기보존을 위한 OAIS의 참고모형의 수집 등록(Ingest)이나 Migration, 열람 제공 등에 유용하게 활용되는 디지털 포맷 레지스트리는 표 1과 같이 국외 각 기관에서 시도되고 있는데, 그 내용은 다음과 같다.

영국 국가 기록원(The National Archives : TNA)의 경우 2002년 The technical registry PRONOM이라는 장기 보존 지원 디지털 객체를 설명하는 기술정보 데이터베이스 서비스를 개발하여 TNA 홈페이지에서 활용하고 있으며, 그 외 DROID (Digital Record Object Identification) 포맷 식별 소프트웨어를 별도로 제공하고 있다[7].

미국의 경우 Global Digital Format Registry (GDFR)이라는 파일 포맷 레지스트리를 구성하기 위한 모델을 개발하여 사용 중에 있으며[1], 하버드 대학과 JSTOR(학술 자료를 보존, 검색, 이용, 구축을 지원하는 비영리 기관) 협업으로 구현한 파일 식별, 오류 검증 프로젝트인 JHOVE를 개발, 구현하여 제공하고 있다[8].

표 1. 각국의 디지털 포맷 레지스트리 구축 사례

국가	구축 사례	설 명
영국	PRONOM	2002년에 영국에서 개발되어 TNA 홈페이지에서 사용되고 있는 서비스로 파일 포맷의 식별, 메타데이터 추출 기능과 마이그레이션을 위한 Metadata 제공 기능 지원
	DROID	TNA에서 개발된 포맷 식별 소프트웨어
미국	GDFR	파일 포맷 레지스트리를 구성하기 위한 모델로 분석 모델, 포맷 모델, 데이터 모델, 구분자 정의 방법 등을 정의
	JHOVE	하버드 대학과 JSTOR (학술 자료를 보존, 검색, 이용, 구축을 지원하는 비영리 기관) 협업으로 구현한 파일 식별, 오류 검증 프로젝트
네덜란드	네덜란드 국립 도서관 KB	e-Depot 시스템 개발 : 기술적 핵심은 IBM사의 DIAS(Digital Information and Archiving System) 전자 출판물을 대상
뉴질랜드	NLNZ (National Library of New Zealand)	"Metadata Standards Framework - Preservation Metadata" 표준을 재정 디지털 객체 보존 메타데이터 추출 툴 개발
	PREMIS	디지털 객체 보존 계획에서 활용 가능한 핵심 보존 메타데이터(Preservation Metadata) 를 위한 데이터 사전

네덜란드의 경우 네델란드 국립 도서관(KB)에서 디지털 포맷 레지스트리를 활용한 e-Depot 시스템 개발을 개발하여 서비스중에 있으며, 뉴질랜드의 NLNZ(National Library of New Zealand)에서는 디지털 객체 보존 메타데이터 추출 툴을 개발하고[9] "Metadata Standards Framework - Preservation Metadata"의 보존 메타데이터 표준을 재정하는 시도가 이어졌다[10].

그 외 디지털 객체 보존 계획에서 활용 가능한 핵심 보존 메타데이터(Preservation Metadata)를 위한 데이터 사전인 PREMIS(Preservation Metadata : Implementation Strategies) 등 디지털 포맷 레지스트리에 관한 많은 연구가 활발하게 시행되고 있는 중이다[4].

그러나 DROID의 예에서 보았듯 기존 포맷 식별 서비스는 대부분 국외에서 구축된 사례로, 국내에서 주로 쓰이는 포맷이나 국내 공공기관에서 쓰이고 있는 레거시 포맷에 대한 고려가 없어 그대로 적용하기에는 큰 어려움이 따른다.

또한 JHOVE와 같이 기존의 포맷 검증 및 특성 추출시스템은 국내 문서의 주류를 이루고 있는 MS 오피스 문서나 국내 공공기관 문서의 대부분을 차지하는 한컴 오피스 문서, 레거시 문서에 대한 고려가 없어 그대로 적용하기에는 큰 어려움이 따른다.

한편 보다 효율적인 서비스를 제공하기 이하여 사용자 입장에서 가장 필요로 하는 기능 중 하나는 원하는 데이터 쉽게 찾는 방법을 제공하는 것이며, 관리자 입장에서는 검색의 키가 되는 데이터(주로 텍스트)의 자동생성 기능을 가장 필요로 한다.

그러나 현재 활용되는 대부분의 기록 관리 시스템에서는 디지털 객체를 저장할 때 저작자나 다른 부가적인 시스템을 활용하여 객체 내에 존재하는 텍스트 기반의 콘텐츠를 추출하며, 일일이 수작업을 통해 해당 객체를 색인 및 검색할 수 있도록 "키워드" 혹은 근래에는 "태그"라는 메타 정보를 첨부하고 있다.

이와 같이 국외 디지털 포맷 레지스트리에 관한 여러 사례를 통해 보았듯 디지털 포맷 레지스트리의 중요성은 더 이상 강조하지 않아도 될 정도이다. 그럼에도 불구하고 현재 국내에서는 디지털 포맷 레지

스트리의 구축 사례를 물론 연구조차 제대로 이루어지고 있지 않은 실정이다. 따라서 본 연구에서 수행한 "디지털 포맷 및 애플리케이션 기술정보은행(DFR) 프로토타입 개발이 국내 디지털 포맷 레지스트리 연구, 구축의 시작점이라고 할 수 있다. 뿐만 아니라 본 연구에서 구현된 디지털 포맷 레지스트리는 국내 실정에 맞는 형태로 개발되었기 때문에 개인이 소유하고 있는 문서 뿐 아니라, 국내 공공기관이 보유중인 문서들의 장기보존에 보다 효과적으로 활용될 수 있다. 이에 본 연구에서 개발된 디지털 포맷 레지스트리를 국내 실정에 맞게 변형된 "기술정보은행"이라 명명하도록 한다.

3. 기술정보 은행의 구성요소

본 연구의 핵심적 목표는 "국내환경에 적합한 XML 기반의 Digital Format Registry"를 구현하는 것이다. 전자기록의 장기보존과 장기 접근성 유지에 근간이 되는 기술정보(technical information)요소, 즉 포맷정보와 소프트웨어정보를 지속적으로 수집, 관리하여 마이그레이션이나 에뮬레이션과 같은 보존 전략을 효율적으로 지원하기 위한 우리나라 실정에 맞는 포맷레지스트리 시스템의 개발이 주된 목적이라 하겠다. 세부적인 내용은 다음과 같다.

- 포맷 및 소프트웨어 기술정보은행의 데이터 및 서비스 모델 연구 및 개발
- 포맷 및 소프트웨어 기술정보은행을 기반으로 한 포맷 식별, 포맷 검증, 포맷 특성 정보 추출, 포맷 텍스트 추출, 포맷 배포 기능의 연구 및 개발
- XML 기반의 기술정보은행(DFR) 프로토타입 개발

3.1 기술정보 은행의 기본 요소

다음 표 2와 같이 기술정보은행과 관련된 기술 요소 중 포맷에 대한 기술정보는 포맷 명, 버전, 유형, 상태, 지적 재산권 및 지원 기관과 같은 일반적인 요소를 포함한다. 또한 외부, 내부 식별자 및 포맷의 기술 문서, 기술적 환경, 메타데이터 등과 같은 기술적인 요소를 포함한다.

또한 기술정보은행의 구성요소는 소프트웨어에 대한 기술정보를 포함하며 그 내용은 다음 표 3과 같다. 소프트웨어에 대한 기술정보는 해당 포맷을 실행할 수 있는 소프트웨어 및 환경에 대한 정보를 제공하기 위해 필요한 구성요소이다. 특히 소트프웨어명, 버전, 유형, 소장여부, 처리 가능한 포맷 외에 소프트웨어 지원에 필요한 소프트웨어 및 하드웨어 요건을 포함한다.

표 2. 포맷에 대한 기술정보

구성요소명	설 명
FormatID	자동 생성되는 관리 번호
Identifier	포맷 등록 시스템내 식별자
FormatName	포맷명
FormatVersion	버전
FormatAliases	다른 포맷명
FormatType	포맷 유형
Status	포맷 상태
IPR	지적재산권
Developer	개발자
Support	포맷 지원/유지 기관
ReleaseDate	포맷 발행일
WithdrawDate	포맷 지원 종료일
FormatDisclosure	포맷 공개 수준
Orientation	Text기반/Binary기반
Related Foramt	관련 포맷
Document	포맷의 기술 문서
ExternalSignature	외부 서명 항목(확장자)
InternalSignature	내부 서명 항목
TechnicalEnvironment	기술적인 환경
Well-formed Note	Well-formed에 관한 설명 혹은 문서 위치
Validity Note	Validity에 관한 설명 문서 위치
Metadata	포맷 메타-데이터
Note	정보주기

3.2 기술정보식별자(TECHi-PUID)

본 내용에서는 기술정보은행의 포맷/소프트웨어의 식별 아이디인 TECHi-PUID에 대하여 설명한다.

본 문서에서 정의하는 TECHi-PUID는 기술정보은행에서 기술 정보를 분류하고 구분하는 유일한 식별자이며 동시에 포맷과 소프트웨어의 대략적인 환경 및 특성을 표현할 수 있도록 정의되었다. 그러나 서로 다른 포맷 혹은 소프트웨어 정보를 구분할 뿐 같은 포맷을 따르거나 디지털 포맷 혹은 소프트웨어

표 3. 소프트웨어에 대한 기술정보

구성요소명	설 명
SoftwareID	내부 시스템의 식별자
Identifier	소프트웨어 상의 외부 식별자
SoftwareName	소프트웨어 이름
SoftwareVersion	소프트웨어 의 버전 정보
SoftwareAlias	소프트웨어 별칭
SoftwareType	소프트웨어 유형
Status	소프트웨어 상태
Location	소장여부 및 소장 위치
ProcessFormat	처리 가능한 포맷
Language	소프트웨어 의해 지원된 언어
Feature	소프트웨어의 일반적인 특징
Image	썸네일 이미지
IPR	지적재산권
Developer	개발자
Corporation	소프트웨어 개발회사
Support	지원/유지 기관
ReleaseDate	소프트웨어 발행일
WithdrawDate	지원 종료일
Related Sofftware	관련 소프트웨어
Document	소프트웨어의 기술 문서
SoftwareRequirement	소프트웨어 지원에 필요한 운영시스템과 기타 S/W 요건
HardwareRequirement	소프트웨어 지원에 필요한 하드웨어 요건
MediaFormat	소프트웨어 저장 매체 유형
Note	정보주기

각각을 구분하기 위한 생성규칙은 가지지 않는다. 이러한 TECHi-PUID는 앞서 언급하였듯 전체 디지털 객체에 대해서 유일하게 존재하는 유일성과 한 번 할당된 식별자는 외부 환경으로부터 영향을 받지 않는 지속성, 또한 장래 개발된 기술에 유연하게 적용되는 유연성, 간결하게 표현되는 간결성 등의 특징을 가지고 있다.

본 문서에서 정의하는 TECHi-PUID는 두 부분으로 구분된다. 첫 번째 PUID 타입, 두 번째는 실제 구분자이다. PUID 타입은 실제 구분자가 지정하고 있는 디지털 포맷의 기술정보의 분류정보나 그룹 정보 등을 나타내는데 이 PUID 구분자는 동일한 PUID 타입에서 유일한 정보이며, PUID 타입과 PUID 구분자가 조합된 형태의 PUID는 유일해야 한다. PUID 구조는 다음 그림 2와 같이 BNF 표기 방식에 따라

```
<puid> ::= <puid_type> '/' <identifier>
<puid_type> ::= <token> | 'x-' <token>
<token> ::= <국가 기록원에 의해 승인되고 정의된 어떤 형태의 PUID 타입>
<identifier> ::= <fragment> | <identifier> <fragment>
<fragment> ::= <digit> | <letter>
<letter> ::= 'a' - 'z'
<digit> ::= '0' - '9'
```

그림 2. PUID 구조에 대한 BMF 표현

표현할 수 있다.

3.3 기술정보은행의 식별기능

기술정보은행의 가장 첫 번째 처리단계는 포맷 식별기능이다. 포맷 식별이란 현재의 포맷이 어떤 포맷인가를 알아내는 기능으로, 단순히 포맷의 확장자 뿐 아니라 포맷 내부 구조를 확인하고 해당 포맷의 시그니처 정보를 찾는 작업이다. 포맷이란 디지털 객체의 내부적 구조 및 인코딩 정보이다. 이 정보를 활용하여 해당 디지털 객체는 사람이 읽거나 해독할 수 있는 형식으로 표시된다.

본 논문에서 제안한 기술정보은행에서는 각 포맷의 BOF, EOF 혹은 스트림의 일정 위치에 존재하는 서명 정보와 해당 디지털 객체의 확장자명을 활용한 디지털 객체처리 기법을 연구, 개발하였으며 일반적인 포맷 뿐 아니라 국내 포맷의 주류를 이루고 있는 마이크로소프트의 MS오피스 및 한컴 오피스, 공공기관의 레거시 포맷인 훈민정음, 아리랑, 하나워드 등의 식별 가능한 포맷들을 고려하였다.

포맷 식별 기능은 기술정보은행에 구축된 파일포맷 데이터베이스에서 얻은 전체 포맷의 식별 정보와 사용자가 전해준 디지털 객체의 비교를 통해 이루어진다. 해당 디지털 객체가 포맷의 고유한 식별 정보인 시그니처와 매칭되면 포맷 식별 결과에 추가하여 이를 결과 값으로 제공하는 구조로 이루어져 있다.

제안된 기술정보은행의 포맷 식별 처리과정은 다음 그림 3과 같다.

3.4 기술정보은행의 검증기능

기술정보은행에 또 다른 주요 기능은 포맷 검증기능이다. 포맷 검증 기능은 포맷이 올바르게 만들어진 포맷인지 확인하는 기능을 제공한다. 디지털 객체는 객체마다 고유한 구조로 구성되어 있다. 올바르게 만

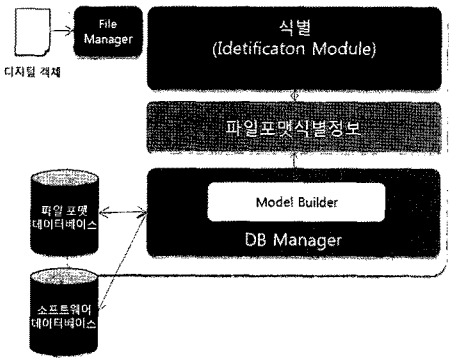


그림 3. 제안 기술정보은행의 포맷 식별처리 과정

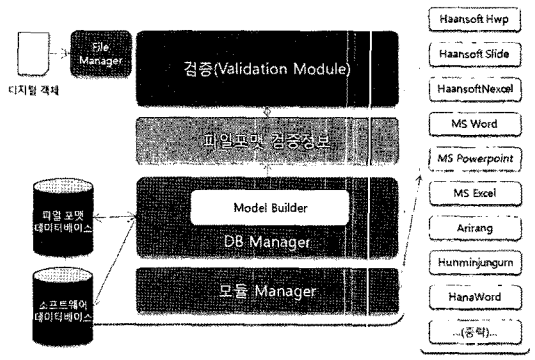


그림 4. 포맷 검증을 위한 처리과정

들어졌는지 확인한다는 의미는 이 객체마다 가지는 고유한 구조에 적합한 형태인지 확인한다는 뜻이다. 검증 기능이 필요한 까닭은 디지털 객체를 보존, 변환 등의 작업을 진행하기에 앞서 해당 객체가 바르게 저장되어 있는 디지털 객체인지를 확인해야할 필요가 있기 때문이다. 바르지 않은 디지털 객체일 경우, 보존의 필요성 자체가 없을 수 있으며, DFR 시스템 내에서 해당 객체를 처리함에 있어 오류를 발생 시킬 수 있다. 포맷 검증 기능은 해당 포맷의 구조에 맞춰 파싱한 후 오류가 없는지를 확인하고 그 결과 값을 사용자에게 전달한다. 포맷 검증기능은 AIFF, ASCII, BYTESTREAM, GIF, HTML, JPEG, JPEG2000, PDF, TIFF, UTF8, WAVE, XML의 일반적인 포맷 외에 마이크로소프트의 MS오피스, (주)한글과컴퓨터의 한컴오피스 및 레거시 포맷인 훈민정음, 아리랑, 하나워드 포맷에 대한 검증을 제공하고 있어 국내 실정에 맞도록 구현되어 있다.

또한 본 기술정보은행에서는 추후 버전이 업그레이드되는 포맷, 혹은 새로이 생성되는 포맷을 위해 관리의 용이성을 가질 수 있도록 검증을 위한 기능을 모듈 형식으로 분리하여 관리하도록 구조화 하였으며 그 내용은 그림 4와 같다.

3.5 기술정보은행의 특성 정보 추출 기능

어떤 디지털 객체의 특성(Characterization)은 전자적 기록물을 저장하는 관점에서는 현재 우리가 어떤 디지털 객체를 가지고 있는가에 대한 메타데이터를 알아내는 것이라고 할 수 있다. 안정적인 기록을 보존하기 위한 분석과 그에 따른 보존 계획을 설계할 때 시작이 되는 곳이기도 하다. 따라서 문서의 중요한 특성을 잘 알아야 기록의 장기 보존을 위한 Ingest 혹은 Migration등의 작업흐름에서 제안 DFR 시스템이 중요한 역할을 할 수 있기 위해서는 디지털 객체에 대한 특성 정보를 제공해 줄 수 있어야 한다.

기술정보은행에 구축된 특성 정보 추출 결과 기능은 특성 정보 추출에 필요한 각각의 모듈을 관리하는 모듈 매니저와 파일 포맷 데이터베이스의 정보를 활용하여 사용자가 원하는 포맷에 대한 특성 정보인 메타데이터 정보를 얻을 수 있다..

포맷의 특성 정보 추출은 AIFF, ASCII, BYTESTREAM, GIF, HTML, JPEG, JPEG2000, PDF, TIFF, UTF8, WAVE, XML의 일반적인 포맷 외에 마이크로소프트의 MS오피스, (주)한글과컴퓨터의 한컴오피스 및 레거시 포맷인 훈민정음, 아리랑, 하나워드 포맷에 대한 특성 정보 추출이 가능하며 다음 표 4는 한컴 워드에 대한 특성 정보 메타데이터의 내용이다.

3.6 기술정보은행의 배포기능

본 "디지털 포맷 및 애플리케이션 기술정보은행(DFR) 프로토타입 개발"에서 개발 구현된 배포기능은 사용자가 특정 디지털 객체에 대한 접근을 위해 필요한 소프트웨어 및 하드웨어등의 정보를 제공하거나 텍스트 정보를 추출하여 전달하여 해당 디지털 객체에 접근이 용이하도록 지원한다. (예: *.dwg -> 필요 소프트웨어 : AutoCad, DWG Viewer, 경로 : www.autodesk.com 등)

장기 보존 관점에서 새로운 기술로 인하여 퇴화된 기술 정보를 이용하여 생성된 디지털 객체의 경우,

표 4. 한컴 워드에 대한 특성 정보 메타데이터

	속 성	내 용	예제 값
한글	format	파일의 포맷	HWP
	version	파일의 포맷 스펙 버전	7.0
	Size	파일의 크기	340
	CreateDate	파일이 생성된 날짜	2008-11-10T12:31
	lastModified	파일이 마지막으로 수정된 날짜	2008-11-10T19:37
	Subject	문서 제목	한글 문서 테스트
	DateTime Modified	수정된 날짜	2008-11-10T19:37:40+09:00
	Author	저작자	김은미
	Doc Summary	문서 요약	한글 문서를 위한 테스트 페이지입니다
	PreView Text	미리보기 텍스트	한글 문서를 위한 테스트 페이지입니다
	reportingModule	현재 파일을 분석하여 결과 값을 전달하는 모듈	HncHwpModule
	App. Version	App 버전	-

접근을 위해 필요한 소프트웨어를 알기 어려우며, 비록 알지라도 해당 소프트웨어를 찾을 수 있는 방법을 알기가 어렵다. 이러한 경우 해당 포맷을 렌더링할 관련 소프트웨어 및 각종 추가적인 정보를 얻을 수 있다.

포맷 배포 기능의 구조는 그림 5와 같이 구성되어 있다.

기술정보은행에 구축된 파일포맷 데이터베이스 및 소프트웨어 데이터베이스를 관리하는 데이터베이스 Manager를 통하여 얻은 파일 포맷 정보와 소프트웨어 정보를 활용하여 사용자가 원하는 디지털 객체가 어떤 포맷인지확인 한 후, 해당 디지털 객체를 렌더링 할 수 있는 소프트웨어정보를 찾아 사용자에게 그 결과를 제공한다.

4. 기술정보은행 시스템의 구현

전체 기술정보은행의 구조는 다음 그림 6과 같다. 먼저 기술정보은행에 필수적인 포맷정보와 소프트웨어 정보를 담고 있는 파일 포맷 데이터베이스와 소프트웨어 데이터베이스가 있으며 이 두 데이터베이스를 관리하여 기술정보은행 서비스와 연동하는 DB Manager가 존재한다.

파일 포맷 데이터베이스와 소프트웨어 데이터베이스를 통해 얻은 정보들을 사용하여 어떤 포맷인지 확인하는 포맷 식별 모듈과, 포맷이 해당 포맷의 고유한 구조에 따라 바르게 작성된 포맷인지 확인하는 포맷 검증 모듈, 그리고 해당 포맷의 중요한 특성 정보를 추출하는 특성 모듈, 포맷의 내용 정보를 얻을 수 있는 텍스트 추출 모듈, 해당 포맷에 종속적이며 포맷

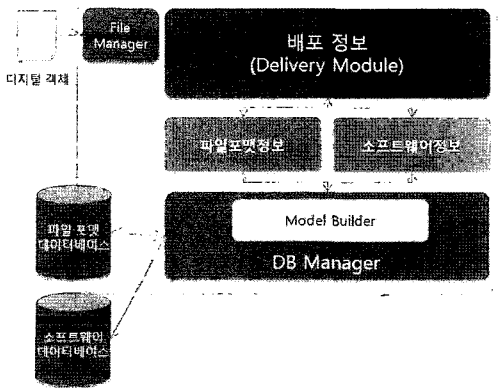


그림 5. 포맷 배포를 위한 제안 시스템의 구조

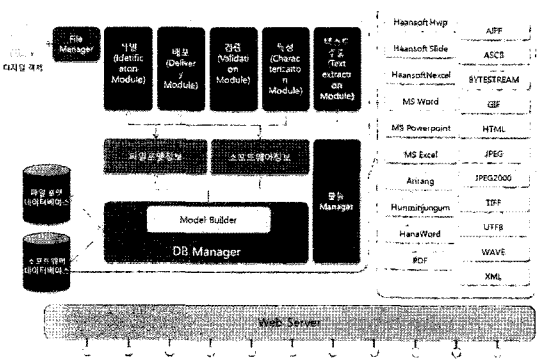


그림 6. 제안 기술정보은행시스템의 전체 구조

의 결과물을 확인할 수 있는 소프트웨어에 대한 정보를 다루고 있는 포맷 배포 정보로 구성되어 있다.

포맷의 검증 및 특성 정보 추출, 텍스트 추출과 같은 결과는 각 포맷의 고유한 구조와 특징에 따라 그 기능 구현이 달라지므로 본 기술정보은행 시스템에서는 해당 기능을 모듈화 하여 새로운 포맷의 추가 및 업그레이드가 용이하도록 구성하였다. 각 모듈을 관리하며, 해당 모듈에 대한 정보를 받아 검증 및 특성 정보, 텍스트 추출 기능과 연동하여 원하는 결과물을 제공하는 모듈 매니저가 존재한다.

각각의 세부 기능들은 계속하여 설명한다.

4.1 포맷 정보 제공 시스템

사용자는 포맷 식별을 통해 얻은 결과나, 기술정보은행에 구축된 전체 포맷 리스트의 결과를 통해 포맷의 결과를 얻고 싶은 포맷의 아이디를 구할 수 있으며, 해당 포맷 아이디를 통해 앞서 언급한 "포맷 정보 제공기능 사용방법"과 같이 포맷 정보를 얻을 수 있다.

그림 7은 기술정보은행에서 공개된 웹서비스 인터페이스를 활용하여 포맷 정보 제공 기능을 통해 얻은 포맷 정보의 결과화면이다.

4.2 포맷 식별 시스템

앞서 언급한 포맷 식별 기능 사용방법을 통해 포맷의 식별결과를 얻을 수 있다. 포맷의 내부 시그니처 정보의 비교를 통해 식별이 이루어지므로 포맷의 외부 시그니처인 확장자에 변화가 있어도 정확한 포맷의 식별이 가능하다.

그림 8은 기술정보은행에 공개되어 있는 웹 인터페이스인 식별 기능을 활용하여 구축한 사례로, 포맷

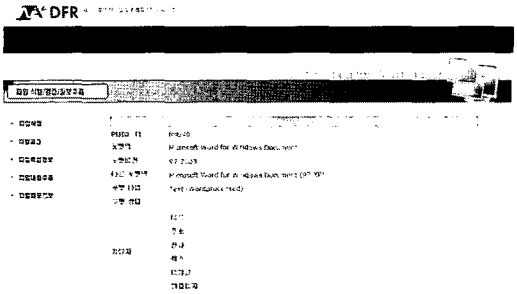


그림 7. 포맷 정보 제공 기능 결과 화면

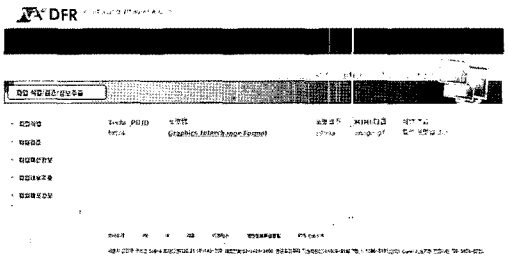


그림 8. 포맷의 식별 정보 결과화면

의 식별 결과를 얻은 화면이다.

4.3 포맷 검증 시스템

포맷의 검증 기능을 활용하여 사용자가 원하는 포맷의 검증 결과를 얻을 수 있다. 그림 9는 기술정보은행에서 공개된 포맷 검증 웹 인터페이스를 활용하여 검증 결과 사용 방법에 따라 구축한 사례로, 해당 포맷의 특성 정보를 얻은 결과 화면이다.

4.4 포맷 특성 추출 시스템

기술정보은행에 구축된 특성 정보 추출 결과 기능은 특성 정보 추출에 필요한 각각의 모듈을 관리하는 모듈 매니저와 파일 포맷 데이터베이스의 정보를 활용하여 사용자가 원하는 포맷에 대한 특성 정보인 메타데이터 정보를 얻을 수 있으며 다양한 특성 추출에 대한 결과는 다음 그림 10과 같다.

4.5 포맷 배포 시스템

포맷의 배포 기능을 활용하여 사용자가 원하는 포맷의 배포 결과를 얻을 수 있다. 그림 11은 기술정보은행에서 공개된 포맷 배포용 웹 인터페이스를 활용하여 배포 결과 사용 방법에 따라 구축한 사례로, 해

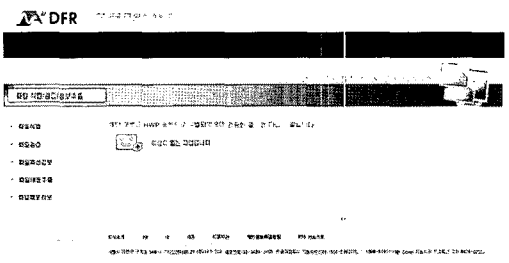


그림 9. 포맷의 검증정보 결과화면

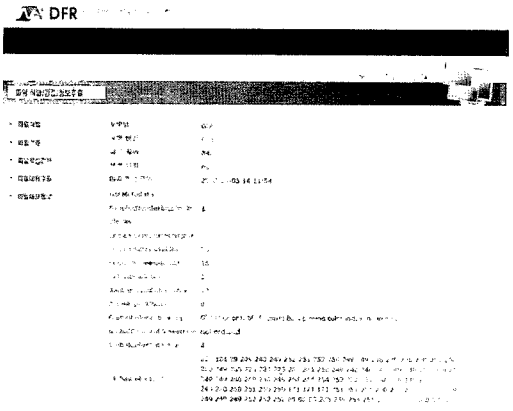


그림 10. 포맷의 특성 정보 추출 결과화면

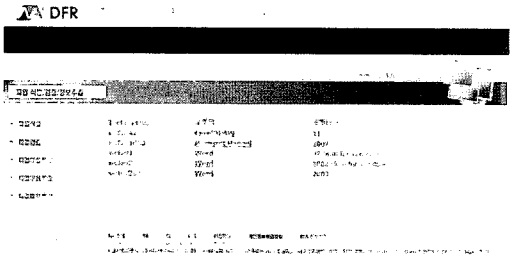


그림 11. 배포 정보 추출 결과화면

당 포맷의 배포 정보를 얻은 결과 화면이다.

5. 결 론

본 연구를 통해 개발된 기술 정보 은행 시스템은 웹서비스 표준에 따라 그 기능을 구현 및 제공함으로 디지털 기록물의 포맷을 처리 및 관리하기 위한 다양한 기능을 네트워크로 연결하여 물리적으로 분산된 어느 곳에서도 활용할 수 있도록 구현하였다.

본 연구의 "디지털 포맷 기술 정보 은행 시스템"을 활용하기 위해서는 기관 차원에서 권위 있고 믿을 만한 곳에서 인증된 정보를 등록, 관리, 유지해야 하며, 구축된 시스템을 중앙 관리를 통해서 사용자들에게 원거리 접근을 가능하게 하고 다양한 형태로 활용 가능할 수 있도록 하는 네크워크형 및 개방형 방식으로 구현되어야 할 것이다.

또한, 기술 정보 은행은 신뢰성 있는 기술 정보를 등록해야 한다. 이를 위해 주기적으로 소프트웨어 관련 업체와 다양한 포맷 정보와 기술 정보를 고유할 수 있도록 해야 하며, 이러한 업무를 위한 채널을 마련하여 수집하여야 한다. 기술 정보 은행과 같은 시스템을 활용하여 기술 정보(포맷 정보, 소프트웨어, 하드웨어 정보 등)을 유지하고 관리하는 목적은 단순한 정보의 보관이 아니라 이러한 정보를 잘 활용해야 하는 중요한 목표가 있다. 따라서 다양한 보존 전략과 사용자의 활용 용도에 맞도록 잘 적용될 수 있어야 한다.

참 고 문 헌

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

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

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

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

[5] 유영수, 2007, "전자기록관리를 위한 포맷등록 시스템 개발 연구," 한국기록관리학회지, 제7권 제1호, 2007.

[6] 설문원, 김연정, 천권주, "ISO/TR18492 의 전자기록 장기 보존 전략," 한국국가기록원구원, 2006.

[7] 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

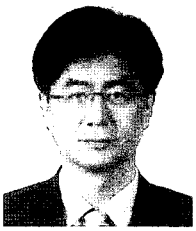
[8] JHOVE, http://hul.harvard.edu/jhove/

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

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

손 원 성

1998년 동국대학교 컴퓨터공학 학사
2000년 동국대학교 컴퓨터공학 석사
2004년 연세대학교 컴퓨터과학 박사
2004년~2006년 Carnegie Mellon University, Post Doc.
2006년~현재 경인교육대학교 컴퓨터교육과 조교수
관심분야 : HCI, 문서처리, 컴퓨터교육

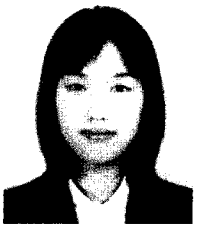


임 순 범

1982년 서울대학교 계산통계학과 (학사)
1983년 한국과학기술원 전산학과 (석사)
1992년 한국과학기술원 전산학과 (박사)
1989년~1992년 (주)휴먼컴퓨터 창업 (연구소장)
1992년~1997년 (주)삼보컴퓨터 프린터개발부 부장
1997년~2001년 건국대학교 컴퓨터과학과 교수
2006년 University of Colorado 방문교수
2002년~기술표준원 전자문서처리위원회 및 ISO/IEC SC34 표준화 위원
2001년~현재 숙명여자대학교 멀티미디어과학과 교수
관심분야 : 컴퓨터 그래픽스, 웹/모바일 멀티미디어 응용, 디지털 방송, 전자출판(폰트, 전자책, XML 문서)

남 동 선

1993년 2월~1997년 2월 용인대학교 전산통계학과 이학사
1997년 3월~1999년 2월 광운대학교 컴퓨터 공학과 공학석사
1999년 3월~2000년 1월 인포텍 시스템 기업연구소 연구원
2000년 2월~현재 (주)한글과컴퓨터 연구개발실 선임연구원 (훈글 개발 등)
2006년 6월~현재 ISO/IEC JTC1 SC34 문서표준 위원회 위원/표준 에디터
2007년 4월~현재 ISO TC46 SC11 보관/기록 관리 표준화 위원회 위원
2009년 4월~현재 TTA PG601 WG6012 디지털교과서 실무반 위원
관심분야 : 문서처리/구조, XML, 전자책, 디지털 멀티미디어 콘텐츠/융합



김 은 미

2000년 3월~2005년 2월 숙명여자대학교 멀티미디어과학과 이학사
2005년 3월~2007년 2월 숙명여자대학교 멀티미디어과학과 이학석사
2007년 2월~현재 (주)한글과컴퓨터 연구개발실 XML 응용기술팀 주임연구원
관심분야 : 문서 표준, 멀티미디어