전자기록 디지털컴포넌트의 포맷과 유효성 검증 기술 연구*

Validation and the Format of the Electronic Record Digital Component Technology Research

이 재 영 (Jae-Young Lee)**
최 주 호 (Joo-Ho Choi)***

목 차

〈초 록〉

전자기록물에 포함된 첨부 파일의 포맷을 이해하지 않은 상태의 전자 기록은 이해할 수 없는 일련의 비트에 지나지 않으므로 기록물을 장기 보존하기 위해서는 포맷의 다양성과 소멸 가능성에 대응할 수 있도록 포맷 정보를 관리하여야 한다. 본 연구에서는 다양한 형식의 전자파일(MS 오피스 계열(PPT, DOC, XLS, PPTX, DOCX, XLSX), HWP, PDF, GIF, JPEG, PNG, TIFF 등)을 육안으로 확인하지 않고, 전자파일 헤더에서 포맷 정보를 추출하여 파일 확장자와 비교하는 포맷검증 도구와 디지털컴포넌트의 유효성을 검사하는 도구를 개발하였다.

주제어: 파일 포맷 검증, 파일 포맷 정보, 기술정보은행, DROID, DFR

<ABSTRACT>

Electronic records are merely series of bits without understanding the formats of content files. There are numerous types of formats and also possibilities of extinction. For long term preservation, it is essential to understand and manage formats. In addition to managing format itself, accurate information on the format needs to be stored for electronic records. In this study, various types of electronic files, without checking with the naked eye, has developed a tool to extract the header information in the format of electronic files with the file extension validation tool to compare format and validate digital component.

Keywords: format validation, digital format registry, PRONOM, DROID, DFR

* 본 연구 논문은 "2011년 행정안전부 국가기록원의 기록물 보존기술 연구개발 사업"의 일환으로 진행한 "차세대 전자기록관리 인프라 응용기술 연구 개발" 과제 결과를 토대로 작성됨.
** (주)세미콘네트웍스(jaeyoung.2ee@esants.co.kr)
*** (주)세미콘네트웍스(iq2chun@esants.co.kr)
■ 접수일: 2012년 11월 14일 ■ 최초심사일: 2012년 11월 26일 ■ 게재확정일: 2012년 12월 20일

1. 서 론

1.1 연구의 필요성과 목적

현행 전자파일 검증 절차는 입수기록접수, 품질확인, 보존패키지 생성, 기술정보 생성, 업데이트 순으로 진행되며 검수 진행 시 오타 검사나 항목 내용의 실제적 정합성 검수는 자동화된 검증 도구에 의해 처리되지 않아 객관적인 검증 체계에 대한 연구가 절실히 요구되고 있다.

따라서 객관적 검증이 가능한 자동화 도구를 개발하여 전자기록물에 대한 신뢰성을 확보하고 기록물 이관 작업의 효율성 확보가 필요하다. 본 연구에서는 2015년 대량으로 인수되는 전자기록물에 대한 포맷 자동 검증 방안을 수립하고 테스트베드 도구를 개발하여 그 적용 가능성을 모색하고자 하였다.

1.2 연구의 범위와 방법

본 연구에서는 다양한 형식의 전자기록물 포맷에 대한 자동 검증을 수행하기 위해 국내외 포맷 식별 및 검증 기술에 대한 선진 사례를 비교 연구하고 개발을 위한 시사점을 도출한 후 이를 토대로 포맷 식별 및 검증 프로세스 수립과 함께 도구를 개발하여 테스트를 함으로써 향후 국가기록원의 전자기록물 이관 업무에 적용하는 기반을 갖추고자 한다.

영구기록관리시스템으로 이관되는 기록에는 기록물철과 기록물건이 있으나, 본 연구에서는 기록물건의 본체에 해당되는 전자 파일의 포맷을 검증하기 위한 각종 요소 기술을 분석하고 검증하는 기능 및 체계를 수립하고, 또한, 향후 더욱 다양해질 포맷 종류를 고려하여 다양한 포맷을 수용할 수 있는 유연성 있는 검증 체계로서 선진 사례에서 도출한 아키텍처를 제시하고 포맷 식별 및 검증 도구 개발시 활용할 수 있는 요건을 개발하여 전체적인 포맷 식별에 대한 방법을 연구하였다.

전자기록 디지털컴포넌트 검증 기술을 연구하기 위하여 장기보존포맷을 비롯하여 전자기록물의 영구보존을 위한 국가기록원 규격[1] 과 장기보존을 위한 메타데이터 규격[2] 및 전자기록물 검증 관련 해외 규격을 검토 분석하였다.

파일 포맷이 식별되고 확장자와 포맷 확장자가 일치하는지 검증을 실시하여 정상적으로 판명된 파일일지라도 비트스트림의 일부가 손상되었거나 암호화되어 사용자가 파일을 열어 볼 수 없는 경우가 존재한다. 이러한 파일은 보존된다 할지라도 내용을 볼 수 없기 때문에 기록이용의 가치가 없다고 할 수 있다. 따라서 기록물의 검수 과정 중에 이러한 파일을 구분할 수 있는 향상된 파일 유효성 검증 기능이 필요하다.

국내외 선행 연구에서 개발된 유효성 검증 도구를 살펴보면 마이크로소프트 오피스 파일에 대해서는 최근에 제한적이나 유효성을 검증할 수 있는 도구가 마이크로소프트에 의해 제공되고 있다. 그러나 이 도구는 윈도우 운영체제에서만 실행되기 때문에 환경 구성이 매우 제약적이라는 큰 문제가 있다. 또한 소스가 공개되

1) NAK-TS 1-2 2008 기록관리시스템과 영구기록관리시스템간 데이터 연계규격.
2) 기록관리시스템 데이터연계 기술규격 제1부 업무관리시스템과의 연계.
기록관리시스템 데이터연계 기술규격 제3부 기능분류시스템과의 연계.

어 있지 않기 때문에 수정 개발이 불가능하고 계속적인 개발 여부도 불투명한 실정이다.[3]

이러한 실정에서 파일의 유효성을 검사할 수 있는 유일한 방법은 육안검수라 할 수 있으나 대량의 기록물을 인수하고 영구적으로 관리해야 하는 영구기록관리시스템에서 육안검수는 현실성이 없는 방법이다.

유효성 검증 도구는 파일의 포맷을 재생할 수 있는 환경에서 파일을 열어 인간이 이해할 수 있는 형식으로 내용을 표시할 수 있느냐를 검증하는 것이다. 이러한 유효성 검증 도구는 다양한 포맷을 지원해야 하며, 그 포맷을 생산한 응용프로그램을 갖추지 않아도 검증이 가능하여야 한다. 또한 운영 플랫폼은 상호운용성을 갖추어야 하며 대량 파일을 일괄 처리할 수 있는 인터페이스와 사용자가 개별적으로 실행할 수 있는 인터페이스(가능한 GUI)를 제공하여야 한다.

2. 국내외 파일포맷 관련 기술 개발 현황

2.1 해외 포맷 레지스트리 기술 동향

2.1.1 포맷 검증 연구 방향

영국 국가기록원, 하버드 대학 도서관을 비롯한 미국과 유럽의 기록보존소와 도서관들은 데이터 포맷을 전자 기록 내에 추상화된 정보가 어떻게 구조화되고 암호화되어 있는지를 설명하는 기술적인 메타 정보[4]로 보고, 포맷을 이해하지 않고는 전자 기록은 이해할 수 없는 일련의 비트에 지나지 않는다는 인식 하에 특정 포맷보다 오랜 시간 기록의 생존을 보장해야 하는 장기 보존 전략의 수립과 시스템 구축에서 포맷의 다양성과 소멸 가능성에 대응하기 위한 연구를 진행하였다.

전자 기록의 보존을 연구해온 연구자들은 포맷 식별과 검증 시 포맷의 다양성으로 인하여 문제에 봉착하게 되며 포맷에 대한 정확한 정보 결여, 불충분한 검증 수단과 포맷 정보 유지관리의 어려움을 지적하였으며 포맷 검증 단계에서 발생하는 문제점에 대하여 〈표 1〉에 정리하였다.

2.1.2 포맷 레지스트리 연구 방향

이처럼 포맷 검증 시 당면한 문제를 해결하기 위해 포맷 정보를 전문적으로 관리하는 포맷 레지스트르의 필요성이 제기된 후 포맷 레지스트리 개발에 연구가 집중되고 있다.

포맷 레지스트리는 각 포맷에 대한 기술 정보는 물론 그 포맷을 생산한 소프트웨어 및 변환에 필요한 정보, 소멸 위험에 놓인 포맷에 대한 정보 등 포맷별로 포괄적이고 전문적인 정보를 관리하는 데이터베이스에 해당하며, 실제 파일의 포맷 식별과 검증은 포맷 레지스트리에서 관리하는 포맷 정보를 이용하되 포맷 레지스트리와 독립적으로 존재하는 별도의 검증 도구가 수행하는 아키텍처를 채택하고 있다.

---

3) Microsoft Office File Format Documents. 〈http://msdn.microsoft.com/en-us/library/cc313105(office.12).aspx〉.

4) MIT 도서관과 HP가 개발 후 현재 비영리 기록관리 연구기관인 DuraSpace의 지원으로 계속적으로 업그레이드 되고 있음. 〈https://wiki.duraspace.org/display/DSPACE/Home〉.

〈표 1〉 포맷 검증 문제점(Problems in validating formats)

항 목	문 제 점
포맷의 다양성	시스템마다 제각기 규정한 포맷명 또는 ID로 포맷을 식별하고 있음. 예를 들면, PDF라고 일반적으로 식별하는 포맷에는 버전 1.3부터 1.6까지 존재하며, PDF/X, PDF/A, Tagged PDF 등 세부 포맷이 존재하며, TIFF 역시 TIFF/EP, TIFF/IT의 세부 포맷이 존재하나 일반적인 포맷명으로는 구분이 안됨.
정확한 포맷 정보 부재	모호한 포맷을 규명하기 위해서 추가적인 정보, 즉 포맷 표준 문서 등을 활용할 수 있는 수단이 필요함.
불충분한 검증 수단	포맷을 구분하는 수단으로 파일 확장자는 불충분함.
포맷 정보 유지관리	포맷 정보의 항목 추가와 변경이 쉬운 저장 방법이 요구됨. 오픈 소스 디지털 리포지토리 중 하나인 DSpace[5]는 포맷 정보를 RDBMS에 저장하였으나 확장성을 위하여 2008년 Release 1.5부터 외부 레지스트리와 연계하여 식별하도록 기능을 추가한 사례가 있음.

대표적인 포맷 레지스트리로서 2002년 영국 국가기록원에 의해 PRONOM이 구축되고 2003년 하버드 대학 도서관의 GDFR(Global Digital Format Registry)이 개발되었다. 이후 포맷 레지스트리는 전자기록의 보존과 시스템 간 상호 호환성을 확보하기 위하여 각종 포맷에 대한 명료한 정의와 식별 정보를 관리하고 제공하기 위하여 개선되어 왔다. 특히 영국 국가기록원의 PRONOM은 기구축 운영중인 디지털 아카이브 및 관련 연구에 가장 많이 활용되고 있다.

2.1.3 검증 아키텍처 연구 방향

외부 포맷 레지스트리와 연계하여 포맷 정보를 얻고 이를 기준으로 식별하는 체계를 의미하며 〈그림 1〉은 오픈 소스 디지털 리포지토리인 DSpace의 포맷 검증 모델로서 외부 포맷 레지스트리인 PRONOM과 연계하고 있다.

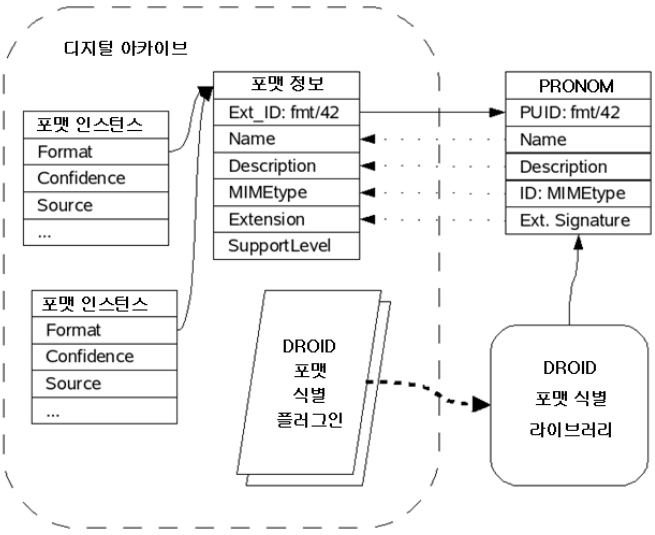
포맷 레지스트리와 별도의 식별 도구를 연계하는 아키텍처는 특정 포맷 레지스트리만 연계하지 않고 선택적으로 포맷 정보를 입수하는 유연성을 갖추고 있으며, 이러한 아키텍처의 성공적 사례인 DSpace는 시스템 설치 구성 파일에 내장 포맷 데이터베이스 또는 외부 레지스트리 중 선택하여 연계하고 있다.

외부 레지스트리로서 PRONOM을 활용한 성공적인 사례인 DSpace는 155,000개 파일을 대상으로 내장 데이터베이스를 기반으로 식별했을 때 21개 포맷이 식별되었고 식별되지 못한 파일이 1,020개(0.65%)에 달하였으나, PRONOM 레지스트리와 DROID를 이용한 결과 52개 포맷이 식별된 결과 162개(0.104%)만이 식별되지 않는 큰 개선 효과를 얻을 수 있었다.[6]

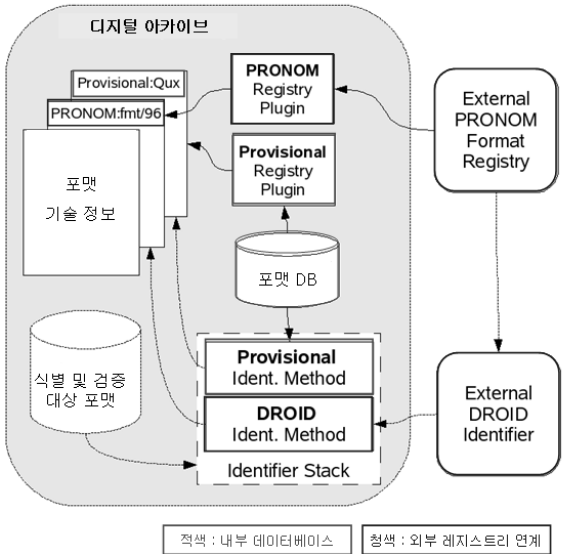
레지스트리 연계를 〈그림 2〉와 같이 레지스트리별 고유한 인터페이스나 프로토콜에 맞추어 개발된 플러그인을 통해 구현하며 식별 및 검증 역시 식별 방법별 또는 레지스트리별로 별

5) MIT 도서관과 HP가 개발 후 현재 비영리 기록관리 연구기관인 DuraSpace의 지원으로 계속적으로 업그레이드 되고 있음. 〈https://wiki.duraspace.org/display/DSPACE/Home〉.

6) Larry Stone. 2008. BitstreamFormat Renovation: DSpace Gets Real Technical Metadata. Open Repositories Conference.



〈그림 1〉 DSpace의 외부 레지스트리(PRONOM) 연계 모델
(A PRONOM linked model for DSpace)



〈그림 2〉 DSpace의 선택적 외부 레지스트리 연계 모델
(A selective PRONOM linked model for DSpace)

도의 플러그인으로 구현함으로써 선택적으로 설치 및 실행할 수 있도록 구성하였다.

플러그인 방식은 영구기록관리시스템의 입수 및 각종 포맷 관련 기능의 프로그램 소스에

주는 영향을 최소화하면서 포맷 레지스트리를 변경할 수 있는 확장성을 제공할 수 있다.

2.2 해외 포맷 레지스트리 연구

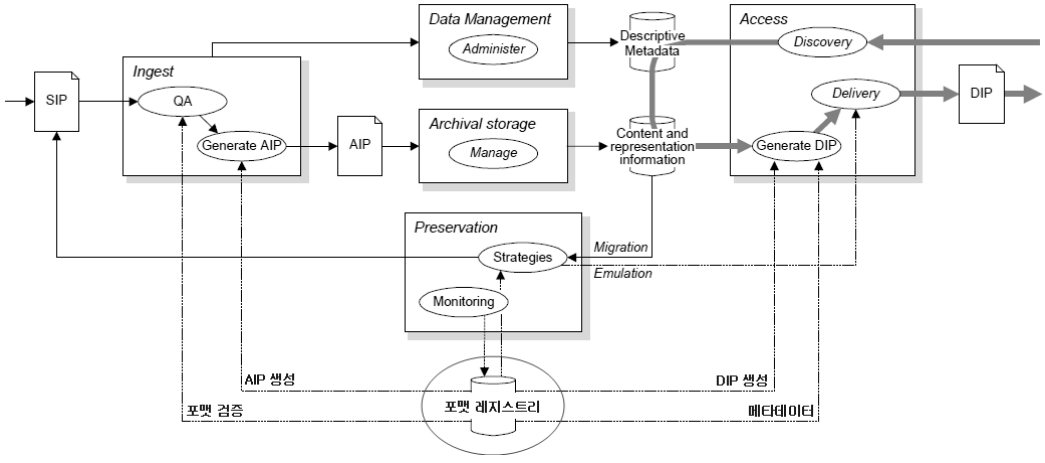
포맷 레지스트리는 영구기록관리시스템이 참조하고 있는 OAIS 모델을 해치지 않고 각 기능과 협력적으로 기능하도록 모델 내에 적절히 위치되어야 한다. 〈그림 3〉은 OAIS 모델의 입수, 보존, 제공 단계에서 포맷 레지스트리를 활용하는 모델이다.[7]

2.2.1 PRONOM

영국 국가기록원(The National Archive: TNA)에서 다양한 포맷의 전자 기록을 장기 보존하는데 활용하기 위하여 2002년 3월 내부 사용을 위해 개발한 후 2004년 2월 무료 온라인 서비스로 공개되었다. 지속적으로 데이터베이스 및 소프트웨어가 업데이트 되고 있으며 2011년 9월 현재 약 820개의 포맷 정보가 저장되어 있다.

PRONOM 서비스[8]는 3개 요소로 구성된다. 첫번째는 포맷 관련 데이터베이스이며 두 번째 요소는 DROID(Digital Record Object Identification)로서 PRONOM에서 제공되는 포맷 정보를 이용하여 파일의 포맷을 식별하는 별도의 소프트웨어이다. 마지막으로 세번째 요소는 PUID(Persistent Unique Identifier)로서 PRONOM 레지스트리의 기록을 지속적이고 유일하며 명확히 식별하기 위한 식별 체계이다.

미국 Harvard 대학교 도서관과 NARA(The US National Archives and Records Administration) 그리고 OCLC(Online Computer Library Center)의 협력 하에 추진되어 온 GDFR(Global Digital Format Registry)와 통합하여 UDFR(Unified Digital Format Registry)를 구축하려는 제안이 2009년에 계획되



〈그림 3〉 OAIS 참조 모델 내 포맷 레지스트리(Format registry in the OAIS referencing model)

7) OAIS 참조 모델.
8) 〈http://www.nationalarchives.gov.uk/PRONOM/Default.aspx〉.

어 2012년까지 추진 예정이며, 구축될 UDFR의 포맷 데이터베이스는 PRONOM의 데이터를 기반으로 구축될 예정이다.

2.2.2 GDFR(Global Digital Format Registry)

미국 Digital Library Federation(DLF)가 2003년 기록의 장기 보존에서 의미있고 재현 가능한 컨텐츠의 보존을 가능하게 하는 핵심 요소로서 포맷을 연구하기 위하여 지원한 워크숍과 다양한 데이터 포맷을 기술하기 위하여 펜실베니아 대학교의 John Ockerbloom이 개발한 TOM(Type Object Model)을 이용하는 실험적 시스템인 FRED(Format REgistry Demonstration) 등 포맷에 대한 연구가 활발하게 진행되고 있다.

선행 연구의 연장선에서 하버드대학교 도서관(Harvard University Library)이 2006년 Andrew W. Mellon Foundation의 연구비 지원을 받아 GDFR(Global Digital Format Registry)을 개발하였으며 인력과 기술에서 OCLC(Online Computer Library Center)가 참여하였다.

GDFR 소프트웨어는 GNU Lesser General Public License(LGPL) 하에 사용할 수 있는 공개 소프트웨어이며, 이 중 OCLC가 개발한 소프트웨어는 OCLC Public LIcense 2.0을 따라야 한다.

GDFR 아키텍처는 GDFR 애플리케이션이 작동되는 분산 레지스트리 구조로서, Java, Apache/Tomcat, Berkeley DB XML, Perl 기반의 아키텍처로서 플랫폼에 비의존적이며 모든 정보는 XML 형식으로 표현되며 데이터 모델은 PRONOM 4의 확장형으로 OCLC IWSA/RFA 프레임웍을 채택하고 있다.

2.2.3 UDFR(Unified Digital Format Registry)

PRONOM과 GDFR 및 기타 포맷 레지스트리 프로젝트 성과를 바탕으로 단일 포맷 레지스트리(UDFR)를 구축하기 위한 제안이 2009년 4월 구성되었다.

UDFR 개발 프로젝트는 미국 의회 도서관 연구비 지원을 받아 2011년 1월부터 2012년 1월까지 수행되며 California Digital Library(CDL)의 University of California Curation Center(UC3)가 프로젝트 관리와 개발을 진행하고 있다.

UDFR은 레지스트리 데이터를 검색, 조회 및 다운로드할 수 있는 웹 사용자 인터페이스를 갖추고 API를 제공하며 DROID로 데이터를 내보낼 수 있도록 개발하며 PRONOM 데이터를 이용할 예정이다.

이 프로젝트에는 영국 국가기록원, 하버드 대학교 도서관, 캐나다 국가기록원, 미국 National Archives and Records Administration(NARA), 네덜란드 국립도서관, 미국 의회 도서관, 뉴질랜드 국립도서관, 캘리포니아 디지털 도서관, 독일 국립도서관 등 그동안 포맷 레지스트리 및 관련 연구를 해온 기관들이 임시 위원회를 구성하여 진행하고 있다.

2.3 해외 포맷 식별 및 검증 연구 동향

2.3.1 DROID(Digital Record Object Identification)

PRONOM을 주관하고 있는 영국 TNA에서 개발한 포맷 식별 소프트웨어인 DROID는

오픈 소스로서 누구나 다운로드하여 사용하고 또한 개발에 참여할 수 있는데, PRONOM 데이터베이스에 저장되어 있는 포맷 정보를 XML 형식의 시그니쳐 파일로 다운로드 받아 포맷 식별이 가능하다.

2.3.2 JHOVE2(JSTOR/Harvard Object Validation Environment)[9]

JSTOR와 하버드 대학교 도서관이 협력하여 개발한 포맷 검증 프레임웍인 JHOVE1을 기반으로 오픈 소스 차세대 포맷 식별 프레임웍과 애플리케이션을 목표로 2008년 미국회도서관의 지원으로 California Digital Library, Portico, Stanford University가 협력하여 JHOVE2 프로젝트를 추진한 결과 2011년 4월 최초 버전을 공개하였다. 디지털 객체의 포맷 식별만이 아니라 검증과 추출 및 평가 기능을 갖추었으며 DROID 시그니처 파일이나 파일 확장자를 이용하여 식별하는데 J2SE기반의 Java 애플리케이션으로서 플랫폼 독립적이다.

2.3.3 Fido(Format Identification for Digital Objects)

유럽 연합의 지원을 받아 2003년부터 4년간 수행된 Planets(Preservation and Long-term Access through Networked Services) 프로젝트를 통하여 개발된 포맷 식별 도구로서 DROID의 소스를 바탕으로 개발되었으며 Python과 Jython 두개의 언어로 개발되어 2개 버젼이 존재하는데, PRONOM 데이터베이스로부터 받은 시그니처 파일과 파일 확장자를 이용하여 식별하므로 PRONOM에 저장된 포맷은 모두 식별 가능하나 PRONOM 시그니처 파일을 그대로 사용하지 않고 자체 형식으로 가공하여 이용하고 있다.

2.3.4 Unix File Utility

Unix 계열 운영체제에서 제공되는 유틸리티로서 매직 파일에 저장되어 있는 매직 넘버를 이용하여 식별한 뒤 식별된 포맷을 MIME 형식으로 보고하는데, 식별 과정은 맨 처음 파일 시스템 검사부터 시작하여 식별에 실패하면 다음 검사인 매직 검사, 그래도 실패하면 언어 검사의 순서로 진행된다.

파일 시스템 검사는 stat 시스템 호출 결과를 이용하여 공백 파일인지 특수 파일인지 여부를 검사하여 공백 또는 특수 파일이 아니면 매직 검사를 실시하여 파일 내 특정 위치에 존재하는 특정 데이터가 매직 파일에 저장된 식별자인지를 검사하는데, 앞의 두 가지 검사로 식별되지 않으면 ASCII, ISO-8859-x, ISO 외 8비트 확장 ASCII 세트, UTF-8, UTF-16, EBCDIC인지 검사함으로써 텍스트 파일 여부를 검사하는데 약 2,000여종의 파일 포맷 식별이 가능하다.

2.3.5 FITS(File Information Toolset)

하버드 대학 도서관의 정보 시스템 부서에서 개발한 통합 도구로서 독립적인 포맷 식별 도구라기보다는 타 기관에서 개발한 JHOVE1, Exiftool, 뉴질랜드 국립도서관 메타데이터 추출 도구, DROID 3.0, FFIdent, Unix File Utility를 통합한 복합 도구로서, 포맷 식별만이 아니

9) 〈https://bytebucket.org/jhove2/main/wiki/documents/JHOVE2-functional-requirements-v1_4.pdf〉, p.6.

라 검증과 추출 기능도 갖추고 있으며 자체 개발 도구인 FileInfo와 XmlMetadata도 포함하고 있다.

2.4 기타 포맷 관련 연구 동향

2.4.1 Metadata Extraction Tool

뉴질랜드 국립도서관(National Library of New Zealand)은 기록의 장기 보존을 위한 메타데이터 표준 프레임워크를 수립하고 표준 메타데이터 구조 및 데이터 모델과 함께 다양한 포맷의 파일로부터 메타데이터를 추출하는 도구를 개발하였다.

2003년 최초 개발 후 2007년 오픈 소스로 공개되었고 현재 Version 3을 다운로드할 수 있도록 하였는데, 이 도구는 Java와 XML을 사용하여 개발되었으며 파일로부터 장기 보존 관련 메타데이터를 자동으로 추출하여 XML 형식 파일로 저장한다.

포맷별로 어댑터가 존재하며 BMP, GIF, JPEG, TIFF, MS Word(version 2, 6), Word Perfect, Open Office(version 1), MS Works, MS Excel, MS PowerPoint, PDF, WAV, MP3, BFW, FLAC, HTML and XML, ARC 포맷을 지원하고 이 외 지원하지 않는 포맷일 경우에는 기본 어댑터가 크기, 파일명과 생성일과 같은 데이터를 추출하도록 구성되어 있다.

2.4.2 Dioscuri

네덜란드 국립도서관(Koninklijke Bibliotheek/National Library of the Netherland)과 네덜란드 국립기록원(Nationaal Archief of the Netherlands)이 공동으로 멀티미디어, 데이터베이스, PDF 문서 등 모든 종류의 전자기록을 장기간 활용할 수 있는 지속성 있는 에뮬레이터 개발 프로젝트를 수행하여 2005년부터 2년간 에뮬레이터를 개발하였다.

개발 완료와 함께 2007년 오픈 소스로 제공되었으며 2011년 1월 현재 0.7.0 버전이 공개되어 있는데, Dioscuri는 지속성과 유연성에 대한 요구사항을 만족시키기 위하여 Java로 개발되어 플랫폼에 비의존적이며 모듈화되어 있다.

2.4.3 Sustainability of Digital Formats

미국 의회 도서관(Library of Congress)은 전자기록의 장기 보존에 있어 디지털 컨텐츠 포맷의 중요성을 인식하고 변화하는 포맷에 대응하여 장기 보존 전략을 세우기 위하여 웹 사이트를 통해 포맷에 관한 포괄적인 연구를 수행하고 있다.[10]

현재 입수할 수 있는 포맷 및 재현하기 위한 소프트웨어에 관한 정보를 수집하고 분석 연구를 통하여 장기 보존에 유리한 포맷 및 불리한 포맷을 구분해 내는데 목적을 두고 있다.

2.5 국내 포맷 레지스트리 및 검증 도구 기술 동향

2.5.1 연구 현황

해외에서는 2000년대 초에 이미 전자 기록의 보존에서 포맷의 영향을 인식하고 포맷 레지스트리와 식별 방법에 대한 연구를 시작하여

10) 〈http://www.digitalpreservation.gov/formats/index.shtml〉.

PRONOM과 GDFR이 이미 구축되어 포맷 정보를 관리하고 디지털 객체의 포맷을 식별하는 방법을 연구하여 DROID 등의 주요 소프트웨어를 개발하여 활용하고 있으나 국내에서는 전자 기록 보존에 대한 연구 역사가 짧은 만큼 포맷 레지스트리나 식별, 검증 등에 대한 연구가 전무하였다.

국내외 관련 연구는 2008년 국가기록원의 기록 관리 연구 개발(R&D) 사업을 통해 수행된 "디지털 포맷 및 애플리케이션 기술정보은행(DFR)을 위한 시스템 설계" 사업이 유일하며, 이 연구 성과로서 기술정보은행의 프로토타입이 개발되었으나 국가기록원은 물론 공공기관의 기록보존소에서 기술정보은행의 기능을 활용하고 있지 않은 실정이다.

2.5.2 국내 포맷 레지스트리 구축 현황

국가기록원의 2008년도 기록 관리 연구 개발(R&D) 사업 지원에 의해 수행된 "디지털 포맷 및 애플리케이션 기술정보은행(DFR)을 위한 시스템 설계" 사업 결과로 디지털 기록물의 재현을 위한 기술정보의 수집 및 관리 정보를 위한 요소들을 정의하고 기술정보를 이용하여 디지털 기록물의 장기보존을 위한 기능을 제공하고 있다. 또한 국내 실정에 맞는 디지털 기록물의 보존을 위한 포맷 레지스트리를 설계하고 이를 기술정보은행(DFR: Digital Format Registry)이라고 명명하였다.

1차 연구 이후 보다 다양한 문서 형식을 지원하기 위한 고도화 사업을 수행하여 업무 환경에 가장 빈번히 사용되는 문서 형식으로 2006년, 2008년에 각각 ISO 26300과 ISO 29500이라는 국제 표준 승인된 ODF(Open Document Format) - (추가 15종의 문서 형식)과 OOXML(Office Open XML) - (추가 3종의 문서 형식) 개방형 오피스 문서 형식을 추가 지원하게 되었으며, ISO 28500 WARC라는 웹 기록물 저장 국제 표준 형식에 대한 유효성 검증 기능을 개발하여 지원하고 있다. 또한 국내 실정을 반영하여 MS 오피스, 한컴 오피스, 하나워드, 아리랑, 훈민정음과 같은 레거시 포맷과 국가기록원의 장기보존포맷인 NEO 패키지 파일에 대한 유효성 검증 기능까지 추가하였다.

2.6 국가기록원 기술정보은행 (DFR: Digital Format Registry)

2.6.1 DFR 정보 구성 요소

기술정보은행은 '포맷에 대한 기술 정보'와 '소프트웨어에 대한 기술 정보'로 구성되며 각 정보는 〈표 2〉 기술정보은행 구성 요소와 같은 항목을 포함한다.

2.6.2 DFR 기능 구성 요소

국가기록원의 기술정보은행은 〈표 3〉에서 보는 바와 같이 파일 식별, 파일 검증, 파일 특성 정보, 파일 배포의 4개 기능으로 구성되어 있다.

2.6.3 DFR 프로토타입 시스템

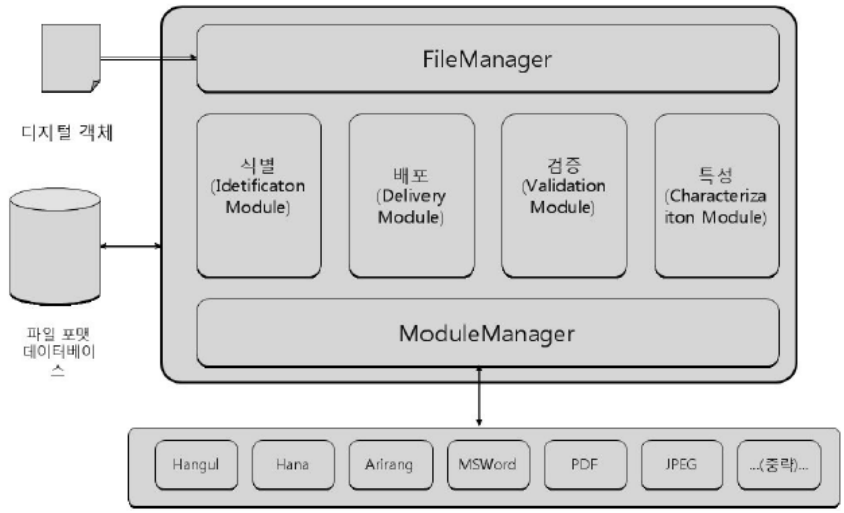
파일 식별, 검증, 특성 정보 추출, 배포 기능을 포맷별로 모듈화하여 포맷의 추가적인 지원에 대비할 수 있도록 설계되었으며, 〈그림 4〉와 같이 식별, 배포, 검증, 특성 기능을 중심으로 플러그인 형식으로 추가된 각 포맷에 대한 모듈들을 모듈 관리자가 관리하는 구조로 설계되었다.

〈표 2〉 기술정보은행 구성 요소(Components of Digital Format Registry)

포맷에 대한 기술 정보	자동 생성되는 관리 번호, 포맷 등록 시스템내 식별자, 포맷명, 버전, 다른 포맷명, 포맷 유형, 포맷 상태, 지적재산권, 개발자, 포맷 지원/유지 기관, 포맷 발행일, 포맷 지원 종료일, 포맷 공개 수준, Text기반/Binary기반, 관련 포맷, 포맷의 기술 문서, 외부 서명 항목(확장자), 내부 서명 항목, 기술적인 환경, Note Well-formed에 관한 설명 혹은 문서 위치, Note Validity에 관한 설명문서 위치, 포맷 메타데이터, 정보주기
소프트웨어에 대한 기술 정보	내부 시스템의 식별자, 소프트웨어 상의 외부 식별자, 소프트웨어 이름, 소프트웨어 의 버전 정보, 소프트웨어 별칭, 소프트웨어 유형, 소프트웨어 상태, 소장여부 및 소장 위치, 처리 가능한 포맷, 소프트웨어 의해 지원된 언어, 소프트웨어의 일반적인 특징, 썸네일 이미지, 지적재산권, 개발자, 소프트웨어 개발회사, 지원/유지 기관, 소프트웨어 발행일, 지원 종료일, 관련 소프트웨어, 소프트웨어의 기술 문서, 소프트웨어 지원에 필요한 운영시스템과 기타 S/W 요건, 소프트웨어 지원에 필요한 하드웨어 요건, 소프트웨어 저장 매체 유형, 정보주기

〈표 3〉 기술정보은행 기능(Functions of Digital Format Registry)

기능	설명
파일 식별	포맷의 확장자 및 바이트스트림 내의 서명정보를 이용하여 포맷을 확인하는 기능. 디지털 기록물에 적절한 처리를 가하기 위해 필수적으로 선행되어야 하는 기능.
파일 검증	디지털 기록물 내에 오류가 있는지 확인하는 기능. 디지털 기록물의 보존, 변환 등 각종 처리 작업에 앞서 손상 여부를 확인함으로써 오류 발생을 줄여 줌.
파일 특성 정보	디지털 기록물의 중요한 특성 정보를 제공하는 기능. 특성 정보는 디지털 장기 보존을 위한 계획 수립 자료로 활용될 수 있음.
파일 배포	특정 디지털 기록물에 대한 접근에 필요한 소프트웨어 및 하드웨어 등의 정보를 제공하는 기능.



〈그림 4〉 기술정보은행 프로토타입 시스템 구성도
(Digital Format Registry prototype system configuration)

가. 식별 모듈

디지털 객체 스트림의 일정 위치에 존재하는 시그니처와 해당 디지털 객체의 확장자명을 활용한 디지털 객체 식별(Identification)기로서, Microsoft Office 바이너리 문서, Open XML 문서(doc, xls, ppt, docx, xlsx, pptx, odp, odt, ods), 오픈 소스 계열의 오피스 문서(StarOffice, OpenOffice 등), 웹 오피스(Google Docx) 등에서 지원되는 ISO 표준 개방형 오피스 문서 표준(Open Document Format) 그리고 한컴 오피스 문서(hwp, hpt, nxl)의 식별 기능을 제공한다.

나. 특성 모듈

디지털 객체의 특성을 알아내는 모듈로서 기록물을 콘텐츠 수준에서 보존하게 하는 해당 포맷의 기본적인 특성 값들인 포맷에 준하거나 포맷 자체가 될 수 있다.

다. 검증 모듈

바르지 않은 디지털 객체일 경우 보존의 필요성 자체가 없을 수 있다. 오류를 포함하는 디지털 객체의 경우 DFR 시스템 내에서 해당 객체를 처리함에 있어 오류를 발생 시킬 수 있기 때문에 보존, 변환 등의 작업 전에 검증을 수행한다.

라. 배포 모듈

해당 디지털 객체를 배포하기 위한 외부 애플리케이션의 정보 및 운영체제, 하드웨어, 애플리케이션 위치 등에 대한 메타데이터를 제공하는데 디지털 객체 중 사용자 위주의 객체 즉, 각종 오피스 문서 객체 등에 대해서는 표준화와 특정 운영체제에 종속적이지 않은 표준 XML 문서, 표준 바이너리 문서 등으로 변환하여 배포를 위해 사용자가 해당 애플리케이션을 활용하지 않고도 해당 객체의 내용을 확인하고 접근 가능하도록 지원한다.

3. 전자기록 디지털컴포넌트 검증 도구 설계

3.1 검증 도구의 필요성

전자기록물의 보존을 위해서는 오랜 시간이 지나 환경이 바뀌었을 때 보존하고 있는 전자기록물을 열어 볼 수 있느냐, 즉 내용 확인이 가능한가 이다. 이때 중요한 부분은 전자기록물을 보존하기 전에 포맷과 실제 파일을 깨어지지 않고, 정상적인 파일이라고 확인하는 것이다. 아무리 좋은 매체로 저장을 한다고 하더라도 이미 보본하기 전에 전자기록물에 문제가 있다면 아무 쓸모가 없어진다. 이러한 부분을 방지하기

사람이 눈으로 직접 확인하는 것은 많은 시간이 소요가 되고 많은 인력이 소요가 된다. 이러한 부분을 해결의 출발점을 제공하기 위해서 전자기록물을 보존하기 전에 인력을 최소화하고 자동적으로 포맷과 문서가 정상적인지 판단할 수 있는 기술적인 필요성이 대두되었다.

검증의 효율성을 더 하기 위해서 두 단계로 나누어서 개발을 하였다. 첫 번째 단계에서는 파일의 포맷을 검증을 하고, 다음 단계에서는 파일이 실제로 유효한지를 검사하는 단계로 진행하였다.

3.2 파일 포맷 검증 도구 설계

본 연구에서 개발한 시스템은 원문 및 별첨 파일의 헤더에서 인식 정보를 추출하여 파일명으로 등록된 확장자와 비교 검증을 수행하였는데 테스트 시스템에서 식별 가능한 포맷은 국가기록원 DFR에 등록된 파일 포맷 중 훈민정음 포맷을 제외한 약 50종의 포맷에 대하여 식별이 가능하다. 즉 HWP, MS 오피스 계열 문서(PPT, XLS, DOC 등), PDF, 이미지 파일 등 약 50여종이며, 훈민정음과 하나 워드프로세서 등 현재 사용되지 않는 애플리케이션 프로그램에 대해서는 지원되지 않고 있다.

또한 파일을 생산한 응용 프로그램에서 기본 포맷이 아닌 타 응용 프로그램 저장 형식으로 저장한 파일의 경우에도 비교 검증이 가능하며 암호화된 파일도 헤더 정보는 인식이 가능하므로 확장자 포맷 비교 검증이 가능하다.

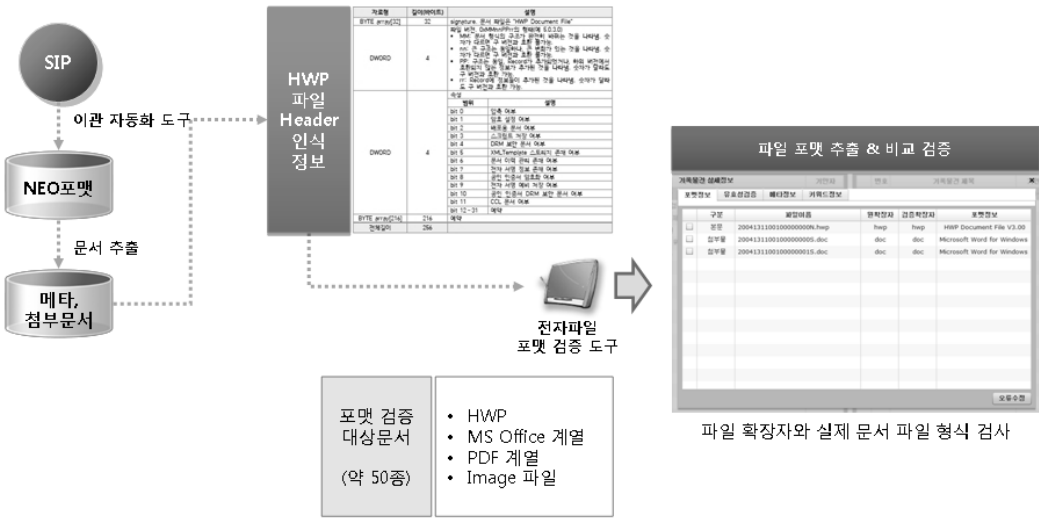
〈그림 5〉에 포맷 식별 및 검증 과정에 대하여 간단한 프로세스를 제시하였다.

한글 문서인 경우는 "HWP Document V3.0", PDF 문서인 경우는 "%PDF-1.4"와 같이 문서가 시작하고, 여기에서 Header부분을 추출하여 확장자와 비교를 비교하여 포맷 유효성을 검증한다.

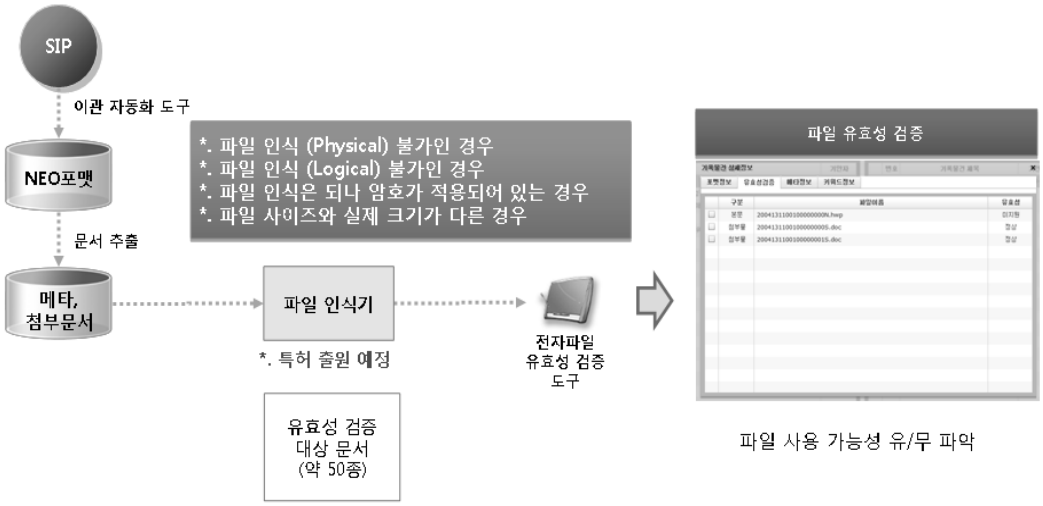
3.3 파일 유효성 검증 도구 설계

본 연구에서 개발한 전자기록 디지털컴포넌트 유효성 검증은 본 연구를 통해 개발한 파일 인식기에 의하여 수행되며 현재 국내 문서 포맷을 제외한 일반적인 파일 포맷에 대하여 검증이 가능하도록 구현되었다. 〈그림 6〉은 파일 유효성 검증 프로세스에 대한 간략화된 설명이다.

본 연구에서는 파일의 유효성을 검증하는 도구를 개발하여 특허 출원을 진행중인데, 파일의 내용을 확인할 수 없는 경우 중에서 가장 빈



〈그림 5〉 포맷 식별 및 검증 과정(Processes for format identification and verification)



〈그림 6〉 파일 유효성 검증 과정(Process for file validity verification)

번히 발생하는 다음 4가지 상태에서 해당 파일의 유효성을 식별하는 도구를 개발하였다. 물리적인 손상으로 파일을 식별할 수 없는 경우, 논리적인 손상으로 파일을 식별할 수 없는 경우, 암호가 적용되어 파일의 내용을 식별할 수 없는 경우 마지막으로 파일 크기와 실제 크기가 다른 경우 등이다.

그러나, 파일을 생산한 응용 프로그램에서는 읽기가 가능하나 파일 내용의 일부가 손상된 경우에는 유효성 판별이 불가하다.

유효성 검증 도구의 개발 가능성을 모색하고 다른 검증 기능과 간섭없이 실행할 수 있는지를 테스트하기 위하여 시스템을 목표로 하였다.

기록물건 정보에서 파일의 위치정보를 파일인식기(SmartCheck v1.0)[11]로 넘겨주면, 파일인식기에서는 문서의 오류를 검사하고, 파일이 유효한지 아니면 이상이 있는 것인지에 대한 결과를 돌려준다. 그 결과를 바탕으로 문서의 유효성 여부를 웹화면에 표시한다.

4. 전자기록 디지털컴포넌트 검증 도구 구현 사례

4.1 도구 개발 환경

본 연구에서는 포맷과 유효성 검증을 위한 시스템을 구현하고자 프레임워크으로는 iBatis 3를 사용하였고, 프로그램밍 언어로는 JAVA와 JSP로 사용하였고, 웹에 표현하기 위해 HTML, CSS를 사용하였다. 기록물철 정보, 기록물건 정보, 첨부물 정보, 파일 포맷 정보, 사용자 정보는 Oracle 10g에 저장하였다.

11) SmartCheck v1.0은 파일의 오류를 검사하는 도구.

〈표 4〉 포맷 및 유효성 검증도구 개발환경 (Development Environment)

분류	개발환경
OS	Windows 2003 server
데이터베이스	Oracle 10g
프레임워크	iBatis 3
웹서버	Tomcat 6.0
개발언어	JAVA, JSP

4.2 파일 포맷 검증 기술 구현 결과

전자기록 디지털컴포넌트 포맷 검증 기능의 사용자 웹 인터페이스는 〈그림 7〉 테스트베드 메인화면에서와 같이 주요 4개 기능별 탭으로 구성되어 있고 각 탭에서 처리과와 생산년도를 선택하며 '파일포맷검증 실행'과 '파일포맷검증 결과' 버튼을 사용하여 선택된 전자기록물에 대한 파일 포맷 검증을 실행한다. 워크스페이스의 좌측은 기록물철에 대한 목록 영역이고, 우측은 해당 기록물철에 대한 기록물건 목록 영영으로서 기능을 실행하거나 결과를 조회하고자 하는 대상 기록물철이나 기록물건을 선택할 수 있다.

'파일포맷검증 결과' 버튼을 클릭하여 〈그림 7〉 파일포맷검증 결과 화면과 같이 검증 결과를 한꺼번에 조회할 수 있는데 결과는 정상과 비정상 목록으로 표시되며 파일이 가지고 있던 확장자(원확장자)와 식별된 확장자(검증확장자)가 표시된다. 검증 결과를 토대로 원확장자와 검증확장자가 불일치되는 기록물철, 기록물건에 대한 파일 목록을 추출하여 보고서를 제공한다.

〈그림 7〉 파일포맷검증 결과 화면(Screenshot for file format validation results)

샘플 데이터는 50개의 포맷 오류 문서는 파일의 확장자를 변경하여 임의로 생성한 40개의 문서와 10개의 인터넷에서 찾은 문서로 검증을 하였다. 300개의 문서 중에서 50개의 포맷 오류 문서를 검증한 결과 모두 정확하게 검증되었다.

4.3 파일 유효성 검증 기술 구현 결과

파일 유효성 검증은 물리적 또는 논리적인 인식 가능 여부, 암호 적용 여부, 물리적인 파일 크기가 파일 헤더 정보와 다른 경우에 대하여 수행되는데, 전자기록 디지털컴포넌트 유효성 검증 기능의 사용자 웹 인터페이스는 〈그림 8〉의 파일 유효성 검증 화면과 같이 구성된다. 전자기록 디지털컴포넌트 포맷 검증과 같이 탭에서 처리과와 생산년도를 선택하며 '파일유효성검증 실행'과 '파일유효성검증 결과' 버튼을 사용하여 선택된 전자기록물에 대한 파일 유효성 검증을 실행한다. 포맷 검증과 마찬가지로 워크스페이스의 좌측은 기록물철에 대한 목록 영역이고, 우측은 해당 기록물철에 대한 기록물건 목록 영영으로서 유효성 검증 기능을 실행하거나 결과를 조회하고자 하는 대상 기록물철이나 기록물건을 선택할 수 있다.

기록물철이나 기록물건을 선택한 뒤 '파일유효성검증 실행' 버튼을 클릭하여 파일 유효성 검증을 실행하면 검증 중인 파일 목록이 상세정보 창에 표시되며 파일 식별이 되지 않는 경우 비정상으로 판정한다. '파일유효성검증 결과' 버튼을 클릭하면 검증 결과를 조회할 수 있는데 결과는 정상과 비정상 목록으로 표시된다.

〈그림 8〉 파일 유효성 검증 결과(Screenshot for file validity verification results)

포맷 오류 문서 50개를 제외한 250개의 문서를 검증하였다. 샘플 데이터로는 40개의 문서는 리눅스 시스템에서 vi로 파일을 열어서 일정한 부분(앞쪽 100byte, 중간 100byte, 끝부분 100byte)을 없애는 방식으로 오류 문서를 만들었고, 나머지 10개의 문서는 인터넷에서 오류 문서를 다운받아서 사용하였다. 검증 결과는 모두 50개의 오류에 대해서 정확하게 검증하였다. 하지만 파일의 오류는 다양한 형태로 존재를 하는데, 테스트베드 시스템에서는 한정적인 데이터를 사용하여 검증할 수밖에 없었다.

4.4 도구의 활용방법

본 연구에서 개발한 시스템은 전자기록물의 디지털컴포넌트를 자동으로 검증하기 위해서 테스트베드 형태로 구현을 한 것이다. 전자결재에서 생산된 기록물 샘플문서를 가지고 기록물의 기본적인 형태만 구현하여 진행하였다. 기록물철, 기록물건, 생산사, 생산부서, 생산일자, 첨부물의 정보를 가지고 시스템을 구현하였고, 시스템을 이용하여 기록물을 등록/삭제하는 기능을 이용하여 검증을 진행할 수 있다.

이번에 개발한 시스템은 테스트베드로서 개발된 것이기 때문에, 국가기록원 CAMS와 같은 시스템에서 사용하기 위해서는 테스트베드의 검증 부분을 어디서나 자유롭게 사용할 수 있는 OpenAPI 형태로 먼저 구현을 하고, 필요한 부분(문서검증 단계)에서 API를 호출하는 방식으로 사용할 수 있다.

5. 결론 및 제언

본 연구에서는 인수 전자기록물에 대한 포맷 검증과 유효성 검증에 대한 자동 검증 방안을 수립하고 테스트 시스템을 개발하여 그 적용 가능성을 모색함으로써 2015년 대량 전자기록물 인수에 대비하고자 하였다.

대량・대용량 기록물 이관을 대비한 CAMS 시스템 개발 사업을 위해 다양한 형식의 전자파일을 육안으로 확인하지 않고 자동화 도구를 사용하여 포맷 및 정합성 등을 검증할 수 있는 체계 및 기술을 연구하였다.

즉, 영구기록관리시스템으로 이관되는 기록에는 기록물철과 기록물건이 있으나 본 연구에서는 기록물건의 본체에 해당되는 전자 파일의 포맷을 검증하기 위한 각종 요소 기술을 분석하고 검증 기능 및 체계를 수립하였으며 향후 더욱 다양해질 포맷 종류를 고려하여 다양한 포맷을 수용할 수 있는 유연성 있는 검증 체계로서 선진 사례에서 도출한 아키텍처를 제시하고 포맷 식별 및 검증 도구 개발 시 활용할 수 있는 요건을 개발하였으며 마지막으로 포맷 식별 방법을 연구하였다.

본 연구를 통해 개발된 '파일포맷검증 도구'와 '파일유효성검증 도구'는 2015년 대량・대용량 기록물 이관을 대비한 CAMS 시스템 개발 사업에서 기록물 검증 단계에 적용 가능하도록 테스트 시스템으로 개발하였으며 차기 사업에 반영하여 성능 및 기능 향상이 가능하도록 구성하였다.

따라서 본 연구를 통해 개발된 "파일포맷검증 도구" 및 "파일유효성검증 도구"를 대량・대용량 기록물 이관 사업의 검증 단계에

적용하여 파일 포맷 및 유효성 자동 검증에 활용할 수 있다. 이를 뒷받침하기 위해서는 더 다양한 문서의 오류 양식에 대해서 테스트를 진행하여야 하는 부분은 여전히 문제점으로 남아 있다.

마지막으로 본 연구를 통해 다양한 포맷 식별 및 검증에 필요한 핵심 기술을 축적하여 체계를 수립하였는 바, 이러한 성과는 향후 영구기록관리시스템 개선 사업에 적용하여 2015년 대량 기록물 인수 시 검수 자동화를 구현할 수 있으며, 기타 기록관리시스템은 물론 기록생산시스템에서도 포맷 식별과 검증 자동화 프로세스를 적용하여 오류 기록물 생산과 보존을 방지하는데 활용할 수 있다.

참 고 문 헌

국가기록원. 2010. 차세대 전자기록관리 인프라 연구 개발. 50.

국가기록원. 2010. NAK/S 7:2010(v1.1) 연구기록관리시스템 기능요건(v1.1).

송병호. 2004. 해외 전자기록물 관련 동향과 시사점. Computer Software & Media Tech, Vol.4. 2004 Sangmyung University.

송병호. 2009. 기록관리시스템의 현황과 전망. 제9회 한국기록학회 학술심포지움, 69-78.

임진희. 2008. 기록관리시스템 기능 요건 표준의 실무적 해석.『기록학연구』, 18: 139-178.

J.H. Lee and J.S. Ahn. 1996. Using N-Grams for Korean Text Retrieval. ACM SIGIR Conference on Research and Development in Information Retrieval, 216-224.

JHOVE2 Team: Functional Requirements, v.1.4, 1-7.

Johan van der Knijff, Carl Wilson. 2011. Evaluation of characterisation tools Part 1: Identification.

Larry Stone. 2008. Bitstream Format Renovation: DSpace Gets Real Technical Metadata. In: Third International Conference on Open Repositories. 1-4.

Medelyan, O. 2005. "Automatic Keyphrase Indexing with a Domain-Specific Thesaurus." Master Thesis. University of Freiburg, Germany.

Medelyan, O., Witten I. H. 2005. "Thesaurus-based index term extraction for agricultural documents." In: Proc. of the 6th Agricultural Ontology Service(AOS) workshop at EFITA/WCCA 2005, Vila Real, Portugal.

Microsoft Office File Format Documents. 〈http://msdn.microsoft.com/en-us/library/cc313105(office.12).aspx〉.