Journal of Korea Multimedia Society Vol. 23, No. 9, September 2020(pp. 1201-1209)
https://doi.org/10.9717/kmms.2020.23.9.1201

전자기록의 장기보존 전략을 위한 의사결정 프로세스 구현

차 현 철†

Implementation of Decision Making Process for Long-Term Preservation Strategy of Electronic Records

Hyun Chul Cha†

ABSTRACT

Based on the risk factor evaluation for the file format, this paper defines the procedures for presenting the long-term preservation plan for that format and the technical information registry necessary for building the system. This is a procedure to perform a risk assessment for the format, evaluate the risk, and select a long-term preservation strategy based on the information registered in the registry and information on the external signature and internal signature of the electronic record. We also reviewed the criteria for selecting appropriate long-term preservation strategies in the process and provided the criteria for adopting each detailed strategy of migration and emulation, which are long-term preservation strategies. And we implemented this process as a long-term preservation decision support system. This system can be used to provide guidelines for the maintenance, management, service and long-term preservation of information resources of electronic records in public institutions such as National Archives of Korea and Libraries.

Key words: Electronic Records, Long-term Preservation, Decision Making Process, Registry

1. 서 론

정보통신의 발전으로 인한 전자기록물의 등장은 기록물 관리 체계에 많은 변화를 가져 왔다. 디지털 콘텐츠는 공간을 초월한 접근성, 배포의 용이성, 그리고 즉시성 등의 강력한 장점을 가진다. 하지만 이것을 정보차원의 한 부분으로 유지·관리하기 위해서는 휘발성과 기술 의존성과 같은 심각한 약점을 지니고 있어 많은 문제점들이 유발된다. 그래서 장기적인 관점에서 정보자원을 유지·관리하여 서비스하고 후대에까지 보존해서 전달해야 하는 도서관과 정보서비스 기관의 측면에서 디지털 콘텐츠의 수집과 보존은 매우 큰 도전 과제이다[1,2].

전자기록물을 안전하게 보존하고 열람·활용하는 과정에는 많은 난관이 있으며 그 이유로는 다음과 같은 세 가지를 생각해 볼 수 있다[3].

먼저, 기술의 발달과 함께 문서 저장 포맷 및 유형이 지속적으로 변화하고 있으며, 문서를 생성하는 애플리케이션 중에서 더 이상 사용하지 않는 애플리케이션이 다수 발생한다는 점을 들 수 있다. 예를 들어, 보석글, 훈민정음, 하나워드 등의 워드프로세서 유형의 프로그램, CAD, MC, MD, Lotus123 등 스프레드

※ Corresponding Author : Hyun Chul Cha, Address: (36040) Dongyangdae-ro 145, Punggi Youngju, Kyoungpook, Korea, TEL : +82-54-630-1088, FAX : +82-54-630-1179 , E-mail : hccha@dyu.ac.kr
Receipt date : Apr. 1, 2020, Revision date : Jul. 13, 2020
Approval date : Jul. 14, 2020
† Dept. of Computer Software, Dongyang University
※ This study was supported by grant from Dong Yang University in 2019.

시트 유형의 프로그램, 웹에디터 등 다양한 애플리케이션 프로그램들이 한 때 사용되다가 지금은 사용이 중단된 상태이다. 현재 가장 많이 사용하는 한글, 엑셀, 파워포인트 등의 애플리케이션도 다양한 버전이 존재하며 버전 간 파일포맷 및 속성의 호환에 문제가 발생하기도 한다.

두 번째로 OS 및 애플리케이션의 변화 등 컴퓨팅 환경의 변화를 들 수 있다. 전자기록 원본 형식을 지원하는 애플리케이션을 보관하고 있지만 운영 환경이 변화하거나 컴퓨팅 환경이 변화할 경우 보관된 애플리케이션을 더 이상 구동할 수 없는 상황이 발생하게 된다. 현재까지 개발된 대부분의 애플리케이션들이 MS 윈도우 운영체제 환경에서 동작하도록 개발되었으나 향후 클라우드 환경에서 구동하게 하는 기술적, 경제적 조치가 어려운 경우에는 보관한 애플리케이션의 동작이 불가능하게 되어 원천파일을 인식하지 못하는 사태가 발생할 것으로 예측된다.

세 번째, 문서 보존포맷으로 변환된 디지털 컴포넌트의 재현 기술의 한계를 들 수 있다. PDF는 시스템 간 이동성이 좋아 배포용으로 광범위하게 사용되고 있으며 특히 PDF/A는 영구보존 문서용으로 채택되고 있다. 우리나라는 공공기록의 문서 보존포맷과 전자문서 보존포맷으로 PDF/A-1을 채택하고 있다. 그러나 PDF/A-1, PDF/A-2, PDF/A-3는 모두 프린트 상태의 정적인 형태로 문서 외형을 재현하는데 그치고 있다. PDF/A에서는 오디오, 비디오, 자바 스크립트 등 동적인 요소들의 사용이 제한되며, 동적 요소가 포함된 한글, 엑셀, 파워포인트 등의 문서를 PDF/A로 변환할 경우 외형만 재현하고 본문에 첨부된 동영상이나 움직임 효과 등이 렌더링 되지 않아 기록물 생산자의 의도나 내용의 맥락을 완벽하게 보존하지 못하는 한계가 발생하고 있다.

그간 전자기록의 장기보존에서 가장 큰 문제로 지적어온 파일포맷은 보석글, 하나 워드, 훈민정음, 한글 워드프로세스 등 국내에서만 사용되던 워드프로세서 관련 파일들이다. 이들 워드프로세서 중 일부가 개발이 중단되면서 기존 응용프로그램에서 작성된 다수의 전자기록에 대한 재현이 불가능해지게 된 영향이 크다. 이러한 워드프로세서 이외에도 국내에서 한정적으로 사용하던 응용프로그램의 개발이 종료된 경우 대체 파일포맷으로 전환이 불가능해지면서 재현에 대한 많은 고민이 되고 있는 것이 현실이다.

다양한 유형의 전자기록물의 안정적 장기보존을 위해서는 적절한 전략이 필요하게 된다. 전자기록물의 파일 포맷, 개발 버전, 구동 S/W 등의 정보가 없거나 부족할 경우 전자기록은 단기간에 독해불능 및 소실 위협에 노출되게 된다. 기술의존도가 높은 전자기록의 적기 보존을 위해서는 파일포맷, 구동 S/W 등의 기초 기술정보에 대한 DB 구축과 위험평가 및 고지 그리고 보존방안의 마련과 실행을 위한 의사결정 프로세스가 반드시 필요하다 하겠다[4,5].

일반적으로 전자기록은 파일 형태로 디스크 또는 다양한 매체에 저장되거나 데이터베이스에 데이터셋 형태로 저장되는데 우리는 파일 형태의 전자기록만을 고려하기로 한다. 파일 형태의 전자기록물을 안전하게 열람, 활용 및 보존하기 위해서는 먼저 해당 전자기록물에 대한 위험이 평가되어야 한다. 전자기록의 위험을 평가하기 위해서는 우선 전자기록에 대한 분석이 필요한데 이때 분석이란 전자기록을 파일로 저장하기 위해 사용한 파일포맷 정보의 파악을 의미한다. 여기에서 파일포맷이란 전자기록을 디스크 또는 매체에 저장하기 위한 규격으로 모든 전자기록은 파일포맷에 맞춰 저장된다. 따라서 전자기록에 대한 위험 평가는 전자기록에 적용된 파일포맷의 위험도를 평가하는 것으로 이해하여야 한다. 포맷 기반의 위험요소 항목에 대한 위험 평가를 수행하여 그 결과를 바탕으로 하여 장기보존 전략이 제시되어야 할 것이다.

본 연구에서는 전자기록 보존방안의 마련과 실행을 위한 의사결정 프로세스를 고안하고 의사 결정을 위한 판단의 근거를 제시하고자 한다. 아울러 고안된 프로세스는 실제 시스템으로 구현한다.

2. 선행연구

2.1 해외 연구 사례 분석

파일포맷 기반의 위험평가에 대한 연구는 주로 해외에서 수행되고 있으며 영국과 호주, 오스트리아 등 여러 국가에서 Table 1에서 보는 바와 같이 전자기록의 장기보존을 위한 전략 수립 프로젝트를 수행하였다[3].

영국 국가기록원의 PRONOM[6,7]은 파일포맷과 구동 S/W에 대한 기술정보 저장소를 구축하는 프로젝트로써 DB 및 S/W가 지속적으로 업데이트 되고

Table 1. Case Study of Foreign Researches

Project	Perform	Main Content
PRONOM	TNA(The National Archive), UK	• File formats and supporting S/W products • Global digital format Registry constructs
AONS II	NLA(National Library of Australia), AU	• Assistance for long-term preservation of electronic records
DipRec	AIT(Austrian Institute Technology), Austria	• Knowledge base system based on semantic web and ontology
PANIC	Univ. of Queensland, AU	• Present new media preservation plan in arts and sciences
SPOT	National Statistical Office, New Zealand	• Propose a long-term preservation strategy for digital materials

있으며 2016년 현재 약 1400여종의 파일 포맷에 대한 기술정보를 확보하여 서비스하고 있다. 호주 국립도서관에서 개발한 AONS Ⅱ[8,9]는 파일포맷에 대한 위험을 평가하는 프로그램으로 위험에 대한 결과만을 제공하며 장기보존 전략에 대해 명시하고 있지는 않다. DipRec[10,11]은 시맨틱 웹/온톨로지 기반의 지식베이스 시스템을 구축하여 전자기록의 위험요소를 정의하고 평가하는 체계를 구축하였다는 특징을 가진다. PANIC[12]은 디지털 기반의 미디어 아트, 과학 데이터, 뉴 미디어 데이터 보존에 방점을 찍고 뉴 미디어 작품에 대한 장기적 접근이 가능하게 하는 방안에 대해 집중하고 있다. SPOT[13]은 뉴질랜드 통계청 등에서 디지털 자료의 보존을 위해 개발한 규격이다.

PRONOM 등을 비롯한 해외의 연구 사례들은 많은 종류의 파일 포맷에 대한 기술정보를 확보하고 서비스를 하고 있지만 국내에서 많이 사용되고 있는 HWP 등의 파일포맷은 여기에 등록되어 있지 않는 등 이들을 우리나라에서 그대로 사용하기에는 불충분한 점을 가지고 있다. 그러므로 국내 상황에 맞는 전자기록물의 장기보존을 위한 연구가 필요하다 하겠다.

2.2 전자기록 장기보존 방안

전자기록의 장기보존 방안으로 가장 광범위하게 사용되는 것은 크게 마이그레이션(migration)과 에뮬레이션(emulation)으로 나누어 볼 수 있다. 마이그레이션이란 하나의 기술로부터 다른 기술로 디지털 객체를 복제 내지 변환하는 방식으로 디지털 객체의 환경보다는 디지털 객체 자체에 초점을 맞춘 장기보존 전략이라 할 수 있다. 여기에 비해 에뮬레이션은 하드웨어 내지 소프트웨어를 흉내 내어 원래의 디지털 객체를 재현하는 방식으로 원래의 컴퓨팅 환경을 재현하는 관계상 전자기록물을 가장 정확한 형태로 장기보존 할 수 있는 전략이라 할 수 있다. 이 두 가지 전략의 특징을 살펴보면 다음과 같다.

먼저, 장기보존 전략으로 마이그레이션이 가지는 가장 큰 장점은 최신의 기술과 환경에서 기록물을 이용할 수 있다는 점을 들 수 있다. 즉 현재의 환경에 변화를 주지 않고 기존의 기록물을 사용할 수 있기 때문에 추가적인 교육이나 훈련이 필요하지 않다. 또한, 데이터를 재사용하거나 자유롭게 편집 활용할 수 있는 형태로 보존할 수 있으며 변환이 쉽고 다양한 도구가 제공된다는 점을 들 수 있다. 해당 파일포맷의 공급자를 비롯한 다양한 관련자로부터 변환도구가 지속적으로 개발 공급되고 있기 때문에 실행에 있어 개발 및 시행을 위한 비용도 낮다. 마이그레이션은 일반적으로 디지털 객체의 지적 내용을 보존하기 위한 신뢰성이 높은 방법으로 특히 페이지 기반의 기록물의 장기보존에 적합하며, 원래의 응용프로그램을 보유할 필요가 없고 마이그레이션을 위한 절차가 간결하게 잘 확립되어 있다는 장점을 가진다[14].

마이그레이션의 가장 큰 단점은 데이터와 속성들의 잠재적인 손실 가능성 때문에 전자기록 장기보존의 주요 이슈인 전자기록의 무결성과 진본성을 손상시킬 수 있다는 점을 꼽을 수 있다. 마이그레이션은 원본의 논리적 구조 혹은 렌더링(rendering) 방법을 다르게 하여 새로 작성하는 방법으로 빈번한 변환의 과정에서 언제든지 데이터 또는 속성의 일부가 변경될 가능성을 가지고 있다. 이를 감안할 때 마이그레이션 된 디지털 객체는 최초의 원본(original)이란 개

념이 아닌 진본(authentic copy)의 개념으로써 다루어져야 하며 진본성을 보장하기 위한 다양한 수단을 강구할 필요가 있다. 또한, 디지털 개체의 특성이 상실될 우려와 함께, 복잡한 레이아웃의 외형이 손상될 가능성도 있다. 디지털 환경이 변경되는 매 시점마다 마이그레이션 업무를 수행해야할 필요가 발생하며 이는 장기적인 마이그레이션 전략의 수립을 어렵게 하는 한 요소가 되고 마이그레이션 전략이 장기화될수록 비용이 점차 증가하게 됨을 의미한다[14].

호주 국립기록청(NAA)에서는 디지털 정보의 장기보존 전략으로 에뮬레이션이 지닌 장점을 원래 데이터 파일포맷의 '외형과 느낌'(look and feel)을 재생산할 수 있다는 점을 들고 있다[15]. 네덜란드의 Digital Preservation Testbed에서는 에뮬레이션 전략이 지닌 강점으로 원본 형태 그대로 디지털 객체를 보존할 수 있으며 저장되는 각 파일포맷 및 각각의 개별 객체에 대해서도 0에 가까운 증분비용(incremental cost)을 제공할 수 있다고 설명하였다. 이와 더불어 에뮬레이션 전략은 보편적으로 적용할 수 있는 단일적이면서도 일관성 있는 보존 전략을 구현할 수 있게 해주며 사양화된 파일포맷에 있는 기록물을 포기할 필요가 없다는 점도 에뮬레이션 전략이 지닌 강점으로 꼽고 있다[16]. 에뮬레이션에 관한 세계적 연구자인 Stewart Granger 역시 디지털 객체의 장기보존 전략으로 에뮬레이션이 다양한 이점을 지닌다고 보고 있는데 우선 원본 디지털 객체를 가장 정확한 형태로 보존이 가능하다는 점과 더불어 마이그레이션과 달리 시간의 흐름에 따라 지속적인 변환 프로세스를 거치지 않아도 된다는 점을 제시하고 있다. 그리고 마이그레이션 전략이 각 디지털 객체에 따라 서로 다른 형태를 띠게 되는 것에 비해, 에뮬레이션은 일관된 보존 전략을 수립할 수 있다는 점 역시 다양한 보존 전략들에 비해 에뮬레이션이 지닌 강점으로 제시하고 있다[17].

에뮬레이션도 단점이 역시 존재하는데, 호주 국립기록청(NAA)에서는 에뮬레이션의 직접적인 도구가 되는 에뮬레이터 소프트웨어의 생산이 고도의 기술력을 필요로 하고 많은 비용이 소요되며, 사유(proprietary) 소프트웨어의 에뮬레이션은 지적 자산과 저작권 문제를 발생시켜 장기보존 전략상의 효율성과 안정성을 손상시킨다는 점을 에뮬레이션 전략의 한계로 제시하였다[15]. 이와 아울러 Stewart Granger는 에뮬레이션 전략이 고도의 기술 및 비용이 소요될 뿐만 아니라 마이그레이션에 비해 기술적으로 고려할 사항이 많기 때문에 실패할 확률이 높다고 보고 있으며 소유권 및 지적 재산권이 존재하는 경우 에뮬레이션 자체가 지적 재산권과 관련된 문제를 야기할 수 있어 보존 전략상의 효율성과 안전성을 침해할 위험성이 있다고 설명하고 있다. 그리고 에뮬레이션의 대상이 되는 시스템과 소프트웨어 환경 역시 다양하게 존재함으로 지속적인 에뮬레이터 개발이 필요한 점 역시 장기보존 전략으로써 에뮬레이션이 지닌 단점 중 하나로 분석하고 있다[17].

장기보존전략으로 마이그레이션을 선정하였을 경우의 다음과 같은 4가지 세부 마이그레이션 방안이 존재한다[4].

- 매체 전환을 통한 마이그레이션
- 메타데이터 캡슐화
- 표준파일포맷으로의 파일포맷 마이그레이션
- 대체파일포맷으로의 파일포맷 마이그레이션

장기보존전략으로 에뮬레이션을 선정하였을 경우의 세부 에뮬레이션 방안은 다음과 같이 3가지 방안이 존재한다[4].

- 애플리케이션 에뮬레이션(Player 또는 Viewer)
- 운영체제 에뮬레이션
- 구동 환경(매체, 디바이스) 에뮬레이션

3. 제안 방안

본 절에서는 장기보존 의사결정을 위한 프로세스를 제안한고 적절한 장기보존 전략 선정을 위한 평가기준을 제시한다.

3.1 장기보존 의사결정 프로세스

전자기록에 대한 장기보존 의사결정을 수행하기 위해서는 우선 전자기록의 파일포맷 정보를 추출하여 포맷 정보 기반의 위험요소 항목에 대한 위험평가를 수행한 후 그 결과를 바탕으로 포맷 기반의 전자기록에 대한 장기보존 방안을 제시하여야 한다. 장기보존 의사결정을 위한 전체 프로세스를 Fig. 1에 제시하였다.

파일포맷의 특성 정보가 비트스트림 외부에 존재하느냐 내부에 존재하느냐에 따라 외장 시그니처

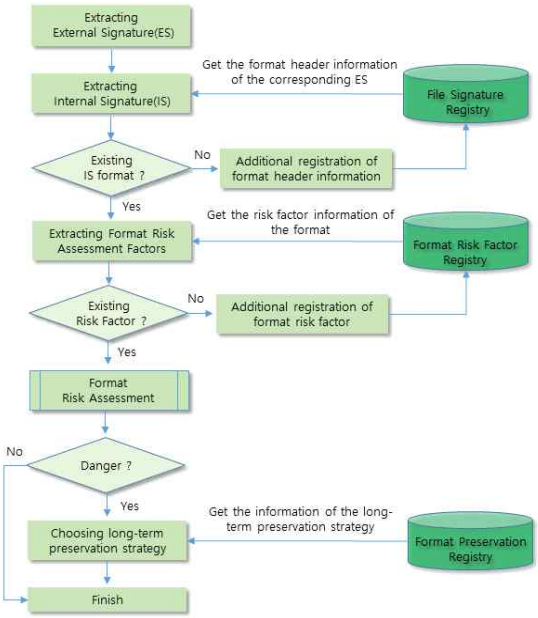

Fig. 1. Long-Term Preservation Decision Making Process.

(External Signature: ES)와 내장 시그니처(Internal Signature: IS)로 구분할 수 있다. 파일포맷과 시그니처와의 관계를 살펴보면, 각 파일포맷은 다수의 내장 시그니처 또는 파일 확장자를 가지며 각 내장 시그니처는 1개 이상의 바이트 시퀀스로 구성된다.

파일 외부에 존재하는 외장 시그니처란 디지털 객체의 비트스트림 바깥에 존재하는 파일포맷 특성 데이터로써 파일 확장자 등이 대표적인 외장 시그니처에 해당한다. 파일 확장자는 특성 값을 얻기는 쉬우나 파일포맷 식별 정보로 불충분한데 그 이유는 첫째, 파일 확장자와 파일포맷의 관계는 1:1의 매핑 관계가 아니기 때문에 특정 파일포맷을 정의하는 유일한 수단이 될 수 없다. 둘째, 파일 확장자는 상세 파일포맷 버전을 표시할 수 없는 문제가 있으며 셋째, 사용자가 확장자를 임의대로 바꿀 수 있기 때문에 신뢰성이 떨어진다는 문제점을 가진다. 따라서 파일 확장자는 파일포맷을 대략적으로 감지하는 용도로 적당하며 그 이상은 신뢰할 수 없다.

파일포맷 내장 시그니처란 디지털 객체의 비트스트림 속의 특정 위치에 존재하는 파일포맷에 대한 특성 값을 말한다. 동일한 파일포맷의 디지털 객체 비트스트림 내에는 일관되게 특정한 구조가 존재하는데 이 구조를 판별해 냄으로써 파일포맷을 판단할 수 있다. 이 구조를 파일포맷 내장 시그니처라고 하며 연속된 16진수 값과 정규식으로 이루어진 1개 이상의 바이트 열로 구성된다.

이 프로세스를 수행하기 위해서는 관련 정보의 저장이 필요하며 본 논문에서는 정보저장을 위한 3개의 레지스트리를 정의한다. 레지스트리란 특정 디지털 정보파일의 구문정보, 의미정보 등을 저장하는 일종의 데이터베이스이다[15]. 먼저 파일의 정확한 포맷 등을 확정하기 위해서는 내부 시그니처를 추출하여야 하며 이를 위해 파일 헤더 위치 정보 등을 저장하는 곳이 필요하다. 이를 File Signature Registry (FSR)이라 부른다. 다음으로 식별된 포맷의 위험 평가를 위해서는 평가 항목별로 점수와 해당 항목에 대한 가중치 정보가 저장되어 있어야 한다. 이러한 정보를 저장하는 레지스트리를 Format Risk Factor Registry(FRFR)이라 한다. 마지막으로 파일 포맷별로 장기보존 전략과 대상 파일 포맷 등의 정보 저장이 요구되며 이를 위해 Format Preservation Registry(FPR)를 필요로 한다. 사용한 레지스트리의 종류와 저장되는 정보를 정리하면 다음과 같다.

- File Signature Registry (FSR): 파일 헤더 정보를 추출하기 위한 헤더 바이너리 정보
- Format Risk Factor Registry (FRFR): 파일포맷별 위험평가 항목에 대한 평가 점수 및 가중치 정보
- Format Preservation Registry (FPR): 파일포맷별 장기보존전략 및 대상 파일포맷 등 마이그레이션 또는 에뮬레이션 대상 정보

장기보존 의사결정 프로세스를 각 단계별로 살펴보면 다음과 같다. 프로세스의 첫 번째 단계에서는 파일의 외장 시그니처(ES)를 추출하는 단계이다. 현재 파일포맷 종류는 약 12만 건 이상 존재하는 것으로 파악되고 있다. 추출한 ES만으로 파일 종류를 유일하게 판정하는 것은 불가능한데 예를 들어 HWP, DAT, SML 등 서로 다른 파일포맷의 파일이 동일한 확장자로 표시되는 경우가 있기 때문이다. 또한 동일한 파일포맷이 여러 개의 확장자를 가지기도 하므로 정확한 포맷 시그니처는 외장 시그니처(ES)가 아닌 내장 시그니처(IS)를 통해서 획득하여야 한다. 이 단계에서는 검증 대상 파일포맷을 문서, 이미지, 오디오, 비디오, XML 등 데이터 파일로 국한할 것인지

아니면 OBJ, EXE 등 실행파일도 포함할 것인지를 결정할 필요가 있다.

두 번째 단계는 내장 시그니처(IS)를 추출하는 단계이다. ES 추출을 통해 대상 파일이 정해지면 외부 확장자를 토대로 IS를 추출할 수 있도록 파일 헤더 위치 정보에 대한 레지스트리 조회가 필요하며 File Signature Registry(FSR)에 확장자 별 파일 헤더의 기술 정보를 관리하게 된다. FSR의 정보를 기반으로 파일 헤더를 추출하여 정확한 파일 포맷명, 버전, 어플리케이션명 등 IS를 추출하는 단계가 두 번째 단계이다. PRONOM에서 제공하는 파일포맷의 종류는 약 1,300개 정도이다. 만약 FSR에 포맷 명에 해당하는 정보가 존재하지 않는다면 해당 정보를 추가 등록한 후 프로세스를 계속하여야 한다.

내장 시그니처를 통해 정확한 파일 포맷명, 버전 및 애플리케이션명 등을 확인하였다면 다음 단계는 해당 파일포맷의 위험평가 항목을 추출하는 단계이다. Format Risk Factor Registry(FRFR)에는 파일포맷별로 위험평가 항목에 대한 점수와 가중치를 관리하고 있으며 여기에서 해당 파일포맷에 대한 점수와 가중치를 추출한다. 만약 FRFR에 해당 파일포맷의 위험평가 항목이 존재하지 않는 경우는 추가 등록하여야 한다.

다음 단계는 FRFR에 등록된 파일포맷의 평가 점수 및 가중치를 기반으로 파일포맷의 위험도를 평가하는 단계이다. [3]에서는 국내에서 주로 사용되는 파일 포맷을 포함하여 전자기록의 장기보존을 위한 위험평가 방식을 제안하고 있으며 본 논문에서는 이 방식에 따라 파일 포맷의 위험도를 평가한다고 가정한다. 이 단계의 평가 결과는 점수로 표시되며 평가 점수는 구간별로 나뉘어 안전(Safe), 보류(Hold), 위험(Danger)의 세 가지 등급 중 한 가지 등급에 속하게 된다. 여기에서 안전 등급은 현행 파일 포맷이 안전하므로 포맷을 그대로 유지하여 사용하는 것에 아무 문제가 없음을 의미한다. 보류 등급은 당장은 아니지만 시간이 경과될 경우 파일 포맷 정보가 소실될 가능성이 존재한다는 것을 의미한다. 이에 비해 위험 등급은 현행 파일 포맷이 장기보존과 활용에 장애를 가지고 있으므로 마이그레이션이나 에뮬레이션 환경 구축 등 장기보존 전략이 필요하다는 것을 뜻한다.

마지막 단계는 위험도 평가 결과 위험 등급을 받은 경우 어떤 장기보존 전략을 취할 것 인지 선택하는 단계이다. Format Preservation Registry(FPR)에는 파일포맷별 장기보존전략과 대상 파일포맷 등 마이그레이션 또는 에뮬레이션 대상 정보가 저장되어 있으며 이 레지스트리로부터 적절한 장기보존 전략을 선정하여 사용자에게 제시하게 된다.

3.2 장기보존 전략을 위한 판단 기준

장기보존전략은 크게 세 가지로 나눠볼 수 있다. 그들은 각각 파일포맷 유지, 마이그레이션(파일포맷 변환), 에뮬레이션(파일포맷 유지)이며 이들 중 적절한 장기보존 전략을 선정하기 위해서는 판단 기준이 있어야 할 것 이다. 이 절에서는 해당 파일포맷에 적절한 장기보존 전략을 선정하기 위한 평가 기준을 제시하며 구체적 기준은 Table 2와 같다.

먼저, 파일포맷 유지 전략을 선정하기 위한 기준은 다음과 같다. 최신의 표준 파일 포맷이거나, 파일포맷 정보가 공개되어 있는 경우, 동종 포맷의 최상위 버전인 경우, 2개 이상의 애플리케이션에서 지원되는 경우 등은 현재의 파일포맷을 유지하더라도 장기보존에 문제가 없다고 판단할 수 있다.

한편 파일포맷 위험도 평가 결과 위험(Danger) 등급인 경우에 마이그레이션이나 에뮬레이션 전략을 선택할 수 있다.

다음은 마이그레이션 전략을 선택하는 기준이다. 마이그레이션에는 매체 전환을 통한 마이그레이션과 메타데이타 캡슐화를 통한 마이그레이션, 표준 파일 포맷으로 마이그레이션, 추천 파일 포맷으로의 마이그레이션으로 나눌 수 있다.

만약 파일 포맷이 특정 구동매체에 종속된 경우에는 매체전환을 통한 마이그레이션을 보존 전략으로 선정하여야 한다. 파일 포맷이 특정 운영체제(벤더)에 종속되거나 혹은 특정 디바이스(벤더)에 종속된 경우에는 메타데이터 캡슐화를 통한 마이그레이션을 보존 전략으로 삼을 수 있다. 여기에서 캡슐화란 디지털 객체와 그 객체로의 접근에 필요한 기타 다른 요소들을 함께 그룹핑 하는 기술을 말한다. 표준파일포맷으로 파일포맷 마이그레이션을 해야 하는 경우는 표준 파일포맷이긴 하나 하위버전이거나 또는 독점 파일 포맷인 경우, 파일 포맷 정보가 공개되어 있는 경우와 애플리케이션의 지원이 종료 버전인 경우가 될 수 있다. 이에 비해 추천 파일포맷으로 파일 포맷 마이레이션 해야 하는 경우는 독점 파일포맷의 하위 버

Table 2. Decision-making criteria for long-term preservation strategies

Decision	Type	Criteria
Maintain file format	-	• The latest standard file formats • File format information is publicly available • The latest version of homogeneous format • Supported by more than one application
Migration	Media transition migration	• Dependent on specific drive media
	Metadata Encapsulation	• Dependent on a specific operating system (vendor) • Dependent on a specific device(vendor)
	Migrating to standard file formats	• Lower version standard file format or proprietary file format • File format information is publicly available • Application support ended version
	Migrate to the recommended file format	• Lower version proprietary file format • If the file format information is private • Application support ended version
Emulation	Application Emulation	• Dependent on a specific application
	OS Emulation	• Dependent on a specific OS(vender)
	Drive Environment Emulation	• Dependent on a specific device(vender) • It is necessary to determine whether the drive environment is emulated according to the importance of the file (cost effective)

전인 경우와 파일포맷 정보가 비공개인 경우 및 애플리케이션 지원이 종료된 버전인 경우가 될 수 있다.

에뮬레이션은 파일포맷 분류가 콘텐츠, 오브젝트 또는 실행 파일포맷인 경우에 해당하며, 에뮬레이션에는 애플리케이션 에뮬레이션, 운영체제 에뮬레이션 및 구동환경 에뮬레이션으로 나눌 수 있다. 여기에서 애플리케이션 에뮬레이션을 하는 경우는 파일포맷이 특정 애플리케이션에 종속된 경우에 선택할 수 있다. 운영체제 에뮬레이션의 평가 기준은 파일포맷이 특정 운영체제(벤더)에 종속된 경우이며 구동 환경 에뮬레이션의 평가 기준은 특정 디바이스(벤더)에 종속된 경우에 고려할 수 있다. 그러나 구동환경 에뮬레이션의 경우 많은 비용이 수반되므로 파일의 중요도에 따라 비용대비 효과를 고려하여 선택 여부를 판단할 필요가 있다.

4. 구 현

4.1 아키텍처 및 프로시져 정의

Fig. 2는 장기보존 의사결정 지원 시스템의 아키텍처를 보여준다. 시스템은 크게 Service 영역, Evaluator 영역, Manager 영역 및 Registry 영역 등 네 개의 영역으로 구성된다. Service 영역은 사용자 인터페이스를 담당하며 서비스 요청을 담당하는 Request 모듈과 결과를 제공하는 Result 모듈로 구성된다. 위험 평가를 위한 Evaluator 영역은 포맷 시그니처를 추출하는 Extract Signature 모듈과 평가를 수행하는 Evaluate Risk 모듈 그리고 장기보존 전략을 제시하는 Decide Preservation 모듈로 구성된다. Manager 영역은 위험평가와 장기보존 전략 선정 등에 필요한 정보를 저장하고 있는 5개의 레지스트리를 각각 관리하기 위한 모듈들로 구성된다. Registry 영역에는 5개의 레지스트리(Format Signature, Risk Index, Risk Weight, Risk Value,

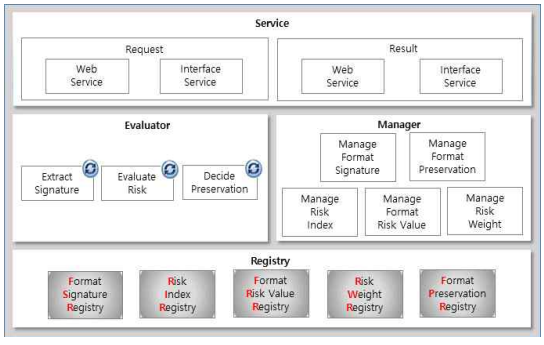

Fig. 2. Module Configuration of Long-term Preservation Decision Support System.

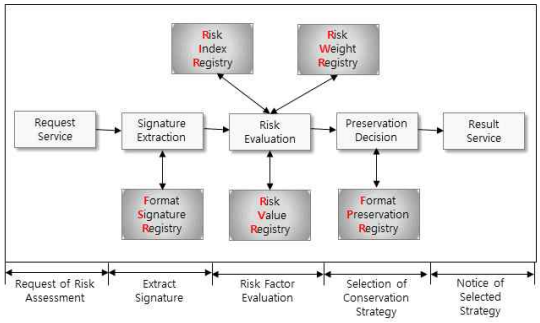

Fig. 3. Procedures of Long-term Preservation Decision Support.

Format Preservation Registry)가 존재한다. 여기에서 Risk Index Registry, Risk Weight Registry, Risk Value Registry 등 세 개의 Registry는 Fig. 1의 Format Risk Factor Registry를 세분화하여 구체화한 것들이다.

장기보존 의사결정 지원시스템의 수행 단계는 Fig. 3에서 볼 수 있는 것처럼 5단계로 구성되며, 위험평가 요청, 시그니처 추출, 위험요소 평가, 보존전략 선정 후 선정 전략을 고지하는 단계로 구성되며 각 단계에서는 앞서 정의된 아키텍처의 서비스(Request Service, Result Service) 모듈과 평가(Signature Extraction, Risk Evaluation, Preservation Decision) 모듈이 동작되도록 절차를 구성하였다.

4.2 시스템 구현

구현된 시스템의 실행결과를 Fig. 4에서 볼 수 있다. 구현된 시스템에서는 전자파일 등록부터 장기보존전략 선정 결과를 한 눈에 파악할 수 있도록 구현하였으며 특정 파일포맷에 대한 상세보존 방안 뿐 아니라 위험평가에 대한 전체 결과를 제공한다. 그러므로 이 시스템을 통해 대상 전자파일에 대한 포맷 정보, 위험 평가 점수 및 장기보존전략 선정 결과를 조회하는 서비스를 제공할 수 있게 된다. 또한 파일 정보부터 포맷 정보, 위험평가 정보, 장기보존전략 선정 정보 등의 상세 정보를 보고서 형태로 출력 가능하도록 구성하였다.

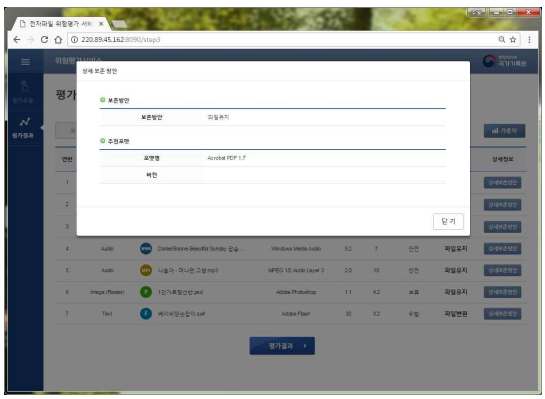

Fig. 4. Implementation of Risk Assessment Decision Support System.

5. 결 론

전자기록물을 안전하게 보존하고 열람·활용하는 과정에는 많은 난관이 있다. 기술의 발달과 함께 문서 저장 포맷 및 유형의 지속적 변화, 문서를 생성하는 애플리케이션 중에서 더 이상 사용하지 않는 애플리케이션의 발생, OS 및 애플리케이션의 변화 등 컴퓨팅 환경의 변화, 문서 보존포맷으로 변환된 디지털 컴포넌트의 재현 기술의 한계 등이 그 이유라 할 것이다.

전자기록물은 기술의존도가 매우 높으며 이들의 적절한 보존을 위해서는 파일포맷, 구동 S/W 등과 같은 기초 기술정보에 대한 수집과 DB 구축, 각 파일 포맷의 위험평가 및 고지 그리고 보존방안의 마련과 실행을 위한 의사결정 프로세스가 반드시 필요하다 하겠다.

본 연구에서 전자기록의 위험도 평가에 관련된 국외 연구 사례들을 살펴본 후 전자기록의 장기보존 방법에 대해 살펴보았다. 다음으로 장기보존 의사결정을 위한 프로세스를 제안하였다. 이 프로세스의 수행을 위해서는 관련 정보의 저장이 필요하며 이를 위해 FSR, FRFR, FPR 등 세 개의 레지스트리를 정의하였다. 장기보존 의사결정 프로세스는 먼저 파일의 외장 시그니처와 내장 시그니처를 추출하여 FSR에 저장된 헤더 정보와 비교하여 정확한 파일포맷 명과 버전 및 애플리케이션 명을 확인하는 식별단계를 수행한다. 그런 다음 FRFR에 등록된 평가점수와 가중치를 기반으로 파일포맷의 위험도를 평가하는 평가단계를 거친다. 평가 결과가 위험 등급으로 평가되었다면 FPR에 저장된 장기보존 전략을 선정하여 사용자에게 제시하게 된다.

우리는 프로세스에서 적절한 장기보존 전략을 선정하기 위한 판단 기준을 검토하였으며 장기보존 전략인 마이그레이션과 에뮬레이션의 각 세부전략을 채택하기 위한 기준들을 제시하였다. 또한 제시한 프

로세스를 구현하기 위한 아키텍처와 프로시져를 정의하였으며 이를 장기보존 의사결정 지원 시스템으로 구현하였다.

본 연구는 국가기록원, 도서관 등의 공공기관에서 전자기록물의 정보자원을 유지, 관리 및 서비스 하고 장기 보존하는데 있어 가이드라인을 제시하는데 사용될 수 있을 것 이다.

REFERENCE

[1] Y. Han, J. Kim, S. Lee, and Y. Lee, "Massive Electronic Record Management System Using iRODS," *Korean Institute of Information Scientists and Engineers Transactions on Computing Practices*, Vol. 16, No. 8, pp. 825-836, 2010.

[2] Y. Jung, H. Yoon, and J. Kim, "A Study on the Preservation Policy for Maintaining the Integrity of Digital Contents," *Journal of Information Management*, Vol. 41, No. 4, pp. 205-226, 2010.

[3] H. Cha and J. Choi, "A Risk Assessment Method for the Long-term Preservation of Electronic Records," *Journal of Korea Multimedia Society*, Vol. 22, No. 1, pp. 79-87, 2019.

[4] Myongji University, *A Study on the Reproduction Technology and the Prototype for the Electronic Records of Administrative Agency*, 11-1312125-000014-01, Korea, 2013.

[5] Myongji University, *Research & Development on Application Technology of the Next Generation Infrastructure for Electronic Records Management*, 11-1311153-000192-01, Korea, 2011.

[6] A. Brown, *Automatic Format Identification Using PRONOM and DROID*, The National Archives 2006.

[7] D. Pearson, "AONS II : Continuing the Trend Towards Preservation Software Nirvana," *Data Analysis and Knowledge Discovery*, Vol. 24, Issue 1, pp. 42-49, 2008.

[8] R. Graf and S. Gordea, "A Risk Analysis of File Formats for Preservation Planning," *Proceedings of the 10th International Conference,* pp. 177-186, 2013.

[9] J. Hunter and S. Choudhury "PANIC: An Integrated Approach to the Preservation of Composite Digital Objects Using Semantic Web Services," *International Journal on Digital Libraries*, Vol. 6, No. 2, pp. 174-183, 2006.

[10] S. Vermaaten, "Identifying Threats to Successful Digital Preservation: The SPOT Model for Risk Assessment," *D-Lib Magazine*, Vol. 18, No. 9/10, 2012. Available online at http://www.dlib.org/dlib/september12/vermaaten/09vermaaten.html.

[11] JCone, *A Study on Long-term Preservation Decision Making Support System Based on Electronic Records Technology Information*, 11-1740083-000035-01, Korea, 2016.

[12] National Archives of Australia, *Digital Recordkeeping : Guidelines for Creating, Managing and Preserving Digital Records*, 2004.

[13] *Emulation : Context and Current Status*, Digital Preservation Testbed White Paper, 2003.

[14] S. Granger, "Emulation as a Digital Preservation Strategy," *D-Lib Maganize,* Vol. 6, No. 10, 2000. Available online at http://www.dlib.org/dlib/october00/granger/10granger.html.

[15] W. Sohn, S. Lim, D. Nam, and E. Kim, "A Study on Digital Format Registry for Digital Objects Preservation in Koreas," *Journal of Korea Multimedia Society*, Vol. 12, No. 10, pp. 1397-1406, 2009.

차 현 철

1989년 경북대학교 통계학과 학사
1993년 경북대학교 컴퓨터공학과 석사
1998년 경북대학교 컴퓨터공학과 박사
1995년~현재 동양대학교 컴퓨터소프트웨어학과 교수
1999년~2000년 (미)애리조나주립대학 방문교수