Journal of Korea Multimedia Society Vol. 22, No. 1, January 2019(pp. 79-87)
https://doi.org/10.9717/kmms.2019.22.1.079

전자기록의 장기보존을 위한 위험평가 방법의 제안

차현철†, 최주호††

A Risk Assessment Method for the Long-term Preservation of Electronic Records

Hyun Chul Cha†, Joo Ho Choi††

ABSTRACT

Appropriate strategies are needed to ensure long-term preservation of various types of electronic records. For proper preservation of electronic records, it is necessary of decision-making processes for risk assessment, notification and implementation of conservation measures. To do this, the task of assessing various risk factors that impede the long-term preservation and utilization of electronic records should be done first. In this study, since electronic records are mostly stored in file form, risk assessment for electronic records of file type is performed. The risk factors required for the risk assessment of the file format are derived, and the algorithms are developed to devise a calculation method of the weighting factor and the risk factor index for evaluating the risk based on the proposed risk factors. In addition, the proposed methods are applied to the file formats used in Korea and risk assessment is performed and the results are analyzed.

Key words: File Format, Risk Assessment, Electronic Records, Long-term Preservation

1. 서 론

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].

다양한 유형의 전자기록물의 안정적 장기보존을 위해서는 적절한 전략이 필요하게 된다. 전자기록물의 파일 포맷, 개발 버전, 구동 S/W 등의 정보가 부재시 전자기록은 단기간에 독해불능 및 소실 위협에 노출되게 된다. 그러므로 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[2-4].

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

먼저, 기술의 발달과 함께 문서 저장 포맷 및 유형

※ 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 : Sep. 20, 2018, Revision date : Nov. 26, 2018
Approval date : Nov. 29, 2018

† Dept. of Computer Engineering, Dongyang University
†† ESan. T. S. Co. Ltd.
(E-mail : E-mail : iq2chun@gmail.com)
※ This research is based on the results of the research conducted as part of the "Research and Development Project for the Archival Technology of the National Archives of Korea in 2016".

이 지속적으로 변화되고 있으며, 문서를 생성하는 애플리케이션 중에서 더 이상 사용하지 않는 애플리케이션이 다수 존재한다는 점을 꼽을 수 있다. 예를 들어, 보석글, 훈민정음, 하나워드 등의 워드프로세서 유형의 프로그램, Lotus123 등 스프레드시트 유형의 프로그램, 그 밖의 CAD, 웹에디터 등 다양한 애플리케이션 프로그램들이 한 때 사용되다가 지금은 사용이 중단된 상태이다. 또한 현재 가장 많이 사용되고 있는 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 있더라도 운영 환경이 변화하거나 컴퓨팅 환경이 변화할 경우 보관된 애플리케이션을 더 이상 구동할 수 없는 상황이 발생하게 된다. 현재까지 개발된 대부분의 애플리케이션은 MS 윈도우 운영환경에서 동작하도록 개발되었으나 향후 클라우드 환경에서 구동하게 하는 기술적, 경제적 조치가 어려운 경우에는 보관한 애플리케이션의 동작이 불가능하게 되어 원천파일을 인식하지 못하는 사태가 발생할 수도 있을 것이다.

세 번째, 문서 보존포맷으로 변환된 디지털 컴포넌트의 재현 기술의 한계를 들 수 있다. 우리나라는 공공기록의 문서 보존포맷과 전자문서 보존포맷으로 PDF/A-1을 채택하고 있다. 그러나 PDF/A-1은 프린트 상태의 정적인 형태로 문서 외형을 재현하는데 그치고 있다. PDF/A에서는 오디오, 비디오, 자바스크립트 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.

그러므로 전자기록물의 장기보존과 활용에 장애가 되는 다양한 위험요소들을 평가하는 작업이 먼저 이루어져야 할 것 이다. 전자기록은 파일 형태로 디스크 또는 다양한 매체에 저장되거나 데이터베이스에 데이터 셋 형태로 저장되는데 본 연구에서는 파일 형태의 전자기록에 대한 위험평가를 수행하고자 한다. 먼저 전자기록의 위험을 평가하기 위해서는 우선 전자기록에 대한 분석이 필요한데 이때 분석이란 전자기록을 파일로 저장하기 위해 사용한 파일포맷 정보의 파악을 의미한다. 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.

본 연구에서는 파일포맷의 위험도 평가를 위해 필요한 위험요소 항목을 도출하고 각 위험요소 항목에 대한 평가 기준을 정립한 후 위험평가를 수행한다.

본 논문의 구성은 다음과 같다. 2장에서는 해외 유사 연구사례와 그 시사점을 분석한다. 3장에서는 파일포맷의 위험도 평가를 위해 필요한 위험요소 항목을 도출하고 각 위험요소 항목에 대한 평가기준을 정립한 후 제안하는 위험도 평가 방식을 설명한다. 또한 제안된 방식에 따라 평가된 위험 평가결과를 다른 연구결과와 비교 분석하였다. 마지막으로 4장에서 결론을 맺는다.

2. 선행연구

2.1 해외 유사 연구 사례 분석

파일포맷 기반의 위험평가에 대한 연구는 해외에서 주로 연구되고 있으며 Table 1의 해외 사례에서 보는 바와 같이 영국과 호주, 오스트리아 등 다양한 국가에서 전자기록의 장기보존을 위한 전략 수립 프로젝트를 수행하였다.

영국 국가기록원에서 다양한 파일포맷의 전자기록을 장기 보존하는데 활용하기 위하여 개발한 PRONOM은 파일포맷과 구동 S/W에 대한 기술정보 저장소를 구축하는 프로젝트로서 DB 및 S/W가 지속적으로 업데이트 되고 있으며 2016년 현재 약 1400여종의 파일 포맷에 대한 기술정보를 확보하여 서비스하고 있다[5,6].

AONSⅡ(Automatic Obsolescence Notification System Ⅱ)는 2006년 소프트웨어 도구의 위험 형식을 관리하기 위하여 호주 국립도서관에서 개발에 착수하였으며 보유한 전자기록의 장기보존을 위해 필요한 행위 결정을 지원한다. AONSⅡ는 파일포맷에 대한 위험을 평가하는 프로그램으로 위험에 대한 결과만을 제공하며 장기보존 전략에 대해 명시하고 있지는 않다[7].

오스트리아 공과대학에서 수행한 DipRec(Digital

Table 1. Case Study of Foreign Studies.

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

Preservation Recommender) 프로젝트는 시맨틱웹/온톨로지 기반의 지식베이스 시스템을 구축하여 전자기록의 위험요소를 정의하고 평가하는 체계를 구축하였다는 특징을 가지며 현재 내부 사용자를 대상으로 서비스를 제공하고 있다[8,9].

PANIC(Preservation webservices Architecture for Newmedia, Interactive Collections and Scientific Data)은 호주 퀸즈랜드 대학에서 현존하는 도구, 서비스, 지원 조직을 발견하여 최적의 보존전략을 수립하기 위하여 개발하였으며 디지털 기반의 미디어 아트, 과학 데이터, 뉴 미디어 데이터 보존에 방점을 찍고 뉴 미디어 작품에 대한 장기적 접근이 가능하게 하는 방안에 대해 집중하고 있다[10].

SPOT(Simple Property Oriented Threat)은 뉴질랜드 통계청 등에서 디지털 자료의 보존을 위해 개발한 규격으로 디지털 자료의 장기보존을 위한 전략으로 6개의 필수 속성 및 해당 속성의 위험요소를 제시하고 있다[11,12].

우리는 앞서 제시한 5개의 유사 사례 프로젝트를 분석하여 다음과 같은 분석 시사점을 도출하였다. PRONOM은 파일포맷 정보 레지스트리 구축의 모범사례로서 외부 시스템과의 연계 서비스 체계 표준을 제시하고 있다. AONSⅡ는 분산형 시스템 아키텍처와 직관적인 사용자 인터페이스를 제공하고 있으며 위험평가 항목에서 가중치 개념을 적용하였다. Dip Rec은 위험속성을 그룹화 하여 그룹별 가중치 개념을 도입한 위험평가 방법을 제시하였고, SPOT은 가장 세부적인 위험평가 규격 설계를 제시하였다.

3. 제안 방법

3.1 위험 평가 기준

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

전자기록에 대한 위험 평가는 전자기록에 적용된 파일포맷의 위험도를 평가하는 것으로 볼 수 있으며, 위험요소는 파일포맷에 대한 위험도를 판단하기 위해 필요한 요소로 정의할 수 있다. 이러한 요소들을 항목별로 나누어 설명한다.

위험도를 판단하기 위해 필요한 첫 번째 항목은 파일포맷의 표준 여부(standardization)이다. 표준은 국제 표준과 국내 표준 그리고 기업 표준으로 구분할 수 있으며, 적용된 파일포맷이 국제 및 국내 표준 파일 포맷인 경우와 기업 표준 또는 특정 소프트웨어에 종속된 독점 파일포맷인지가 중요한 요소로 볼 수 있다.

두 번째 항목은 이해성(understandability)이다. 이해성이란 파일 포맷 정보를 얼마나 이해할 수 있는가를 나타낸다. 이해성을 판단할 수 있는 요소로는 먼저 파일포맷 정보의 공개 여부를 판단해 볼 필요가 있다. 표준 파일 포맷인 경우에는 대부분의 정보가 공개되어 있으므로 위험도가 낮지만 기업 표준 또는

독점 파일포맷이면서 파일포맷 정보가 비공개인 경우에는 위험도가 매우 크다고 볼 수 있다. 또, 원 제조사를 포함한 제품을 지원하는 조직의 수, 해당 응용 프로그램의 버전의 수, 제품의 웹사이트 존재, 기술 문서나 레퍼런스 자료 유무, 출시자와 제품개발자 및 저작권소유자의 정보 유무, 제작일자 유무 등의 요소를 통해 이해성을 판단할 수 있다.

IT 산업의 발전으로 하드웨어 관련 기술이 급속히 진화함에 따라 소프트웨어 역시 빠르게 변화하면서 새로운 기술의 적용을 위해 새로운 파일포맷이 생겨나거나 기존 파일포맷의 버전 업그레이드가 이루어지고 있다. 따라서 세 번째 항목은 지속성(persistency)이다. 독점 파일포맷인 경우 파일포맷의 버전 업그레이드가 진행되며, 표준 파일포맷인 경우에도 지속적인 버전 업그레이드 또는 표준 변경이 이루어지므로 현재 사용 중인 파일포맷인지가 중요하다. 이 지속성의 구체적 요소로는 파일포맷의 버전 수, 파일의 사용 기간 등이 있다.

네 번째 항목은 사용성(availability)이다. 파일포맷의 버전에 대한 평가로서 동일한 파일포맷이라도 버전에 따라 소프트웨어 지원이 종료된 경우가 발생하므로 현재 사용 중인 버전과의 비교를 통한 위험도 평가가 필요하다. 그런데 파일포맷의 버전 정보는 파일의 헤더 정보를 추출하여 확인할 수 있으므로 파일포맷 위험도 평가를 위해서는 파일 헤더에 기록된 내부 시그니처 추출이 반드시 필요하다 할 수 있으며, 내부 시그니처 검증 가능 여부가 네 번째 요소로 사용될 수 있다.

다섯 번째 항목은 소프트웨어 종속성(dependence of software)으로 파일포맷을 지원하는 소프트웨어에 대한 종속성을 의미한다. 표준 파일포맷은 다양한 소프트웨어에서 지원하지만 독점 파일포맷은 특정 소프트웨어에서만 사용이 가능하므로 표준 파일포맷이더라도 단일 소프트웨어에서 지원하거나, 독점 파일포맷인 경우에는 소프트웨어에 대한 종속성이 깊다고 볼 수 있다. 따라서 소프트웨어 종속성이 큰 파일포맷은 지원하는 소프트웨어가 존재하는지 여부를 평가하여야 하고 파일포맷의 버전과는 전혀 다른 소프트웨어의 버전이 존재하므로 소프트웨어 버전에 대한 정확한 정보를 기반으로 위험도 평가가 이루어져야 한다. 소프트웨어 종속성은 지원하는 운영체제의 수, 지원하는 응용 프로그램의 수 등의 요소로 평가할 수 있다.

여섯 번째 항목은 매체 종속성(dependence of media)으로 파일포맷이 저장매체에 종속적인가를 평가하는 것이다. CD나 USB 등 특정 저장 매체에서만 구동되도록 생성된 멀티미디어 파일포맷은 저장 매체를 구동하는 디바이스 지원 여부에 대한 위험도 평가가 필요하다. 매체 종속성은 특정 매체 종속 여부, 지원 매체 숫자, 지원 디바이스 종류, 지원 디바이스 숫자 등의 요소를 통해 평가할 수 있다.

일곱 번째 항목은 재현 가능성(renderability)이다. CAD 파일, 동영상 파일, 3D 파일 등은 소프트웨어 종속적인 경우가 많으며 재현 가능성은 마이그레이션 지원여부, 압축 여부 등의 요소를 통해 평가할 수 있다.

앞에서 제시한 위험요소 항목들 중 이해성, 지속성, 사용성은 정체성(identity)이라는 분류(category)로 묶을 수 있다. 또 소프트웨어 의존성과 매체 종속성은 종속성(dependency)이라는 분류로 묶을 수 있으므로 위에서 제시한 7개의 위험요소 항목은 크게 표준성(standardization), 정체성(identity), 종속성(dependency), 재현가능성(renderability)의 4개 위험도 평가 분류로 나눌 수 있으며 이에 따른 위험요소들을 정리하여 Table 2에 제시하였다.

다음 절에서는 Table 2에 제시된 위험요소 항목을 기반으로 위험도를 평가하기 위한 가중치와 위험요소 지수의 계산 방식을 설명한다.

3.2 위험 평가 방식

Table 2에서 제시된 속성 조합은 해당 조합의 고유한 판단 기준에 따라 YES/NO형태의 판단을 내리게 되며 다차원으로 구성된 속성조합들은 위험평가 단계에서 합산과정을 거치면서 최종적으로 하나의 판단 결과(위험여부)로 수렴된다. 안전과 위험은 서로 대척점에 있으므로 본 논문에서는 특정 파일포맷의 위험 여부를 판단하기 위해 먼저 안전점수를 계산하게 되며, 안전점수를 기준으로 하여 위험도를 평가한다. 이를 위해 각 요소마다 요소 위험지수(Factor Risk Index: FRI)를 정의한다. 요소 위험지수(FRI)는 해당 요소의 고유한 판단 기준에 따라 YES/NO형태의 판단을 내리고 이를 0과 1사이의 지수로 정량화한다. 또한 위험평가 요소들은 각각이 파일포맷의 위험도에 미치는 영향이 다르므로 가중치(weight)를

Table 2. Criteria for risk factor and Risk Assessment Example

Category	Item	Risk Assessment Factor	Description	HWP Format			
				FRI	Asse.	Value	Sum
Standardization (3.0)	Standard (3.0)	International Standard(3.0)	Compliant with international standard file specifications?	0.0	0.0	1.0	1.0
		Korean Standard(2.0)	Compliant with Korea standard file specifications?	0.0	0.0		
		Company Standard(1.0)	Company standard file specifications?	1.0	1.0		
Identity (4.0)	Understandability (2.0)	Format information disclosure(0.4)	Is format information released?	1.0	0.4	2.0	4.0
		Number of product support company(0.2)	The number of organizations that support the product	1.0	0.2		
		Number of application versions(0.2)	Number of versions of the application	1.0	0.2		
		Existing of Website(0.2)	The presence of the product's website.	1.0	0.2		
		Disclosure of technology items(0.2)	Have technical documents, reference materials?	1.0	0.2		
		Publishers information disclosure(0.2)	Publishers information availability	1.0	0.2		
		Creator information disclosure(0.2)	Is the creator's information available?	1.0	0.2		
		Author information disclosure(0.2)	Is the author's information available?	1.0	0.2		
		Presence of production date(0.2)	Is there production date?	1.0	0.2		
	Persistency(1.0)	Number of file format version(0.5)	Number of versions of the file format	1.0	0.5	1.0	
		Existence period(0.5)	Duration of file usage	1.0	0.5		
	Availability(1.0)	Internal signature validation(1.0)	Can internal signature validation?	1.0	1.0	1.0	
Dependency (2.0)	Dependence of S/W(1.0)	Number of supporting OS(0.5)	Number of operating systems supported	0.0	0.5	1.0	2.0
		Number of application program(0.5)	Number of applications to support	0.0	0.5		
	Dependence of media(1.0)	Media dependency(0.4)	Is it dependent on a specific medium?	0.0	0.4	1.0	
		Number of supporting media(0.2)	Number of media supported	0.0	0.2		
		Supported device types(0.2)	Types of devices supported	0.0	0.2		
		Number of devices supported(0.2)	Number of devices supported	0.0	0.2		
Renderability (1.0)	Renderability(1.0)	Migration support(0.5)	Does it support migration?	1.0	0.5	0.5	0.5
		Compression support(0.5)	Is it compressed?	0.0	0.0		
Risk Grade(SAV)				Safety(7.5)			

두어 각 요소의 영향력을 반영할 수 있도록 한다.

위험평가 요소들은 표준성, 정체성, 종속성, 재현가능성 등 모두 4개의 분류로 나눌 수 있었다. 이 중 표준성, 정체성, 재현 가능성은 평가 점수가 높을수록 파일포맷의 안전도가 높아지는 반면 종속성은 점수가 높을수록 파일포맷의 안전도가 낮아지게 된다.

위험 요소 중 상대적으로 중요한 표준성 항목은 다른 위험요소의 결과에 상관없이 절대적 평가를 수행한다. 기업 표준에 대해서는 가중치를 낮게 책정한 것은 특정 시장을 독점하고 있는 제품의 경우에는 해당 제품의 단종이나 인수합병 등이 발생할 때 데이터 마이그레이션이나 복구 방안을 제시했던 현재까지의 경험을 바탕으로 평가정책을 제시하였다.

정체성 분류는 다시 이해성(understandability), 지속성(persistency), 사용성(availability)의 세 항목으로 나눠진다. 이해성, 지속성, 사용성은 점수가 높을수록 상대적으로 위험 요소가 작다는 것을 기본으로 한다. 이 중 사용성을 판단할 수 있는 내부 시그니처에 대한 접근이 불가한 경우에는 다른 위험요소의 점수와 상관없이 안전도 중 이하로 판정하였는데 내부 시그니처에 대한 접근이 불가능한 경우에는 장기보존을 위한 마이그레이션도 불가한 상황이 발생할 수 있다는 가능성을 반영하였다.

종속성에는 소프트웨어 종속성과 매체 종속성이 있으며 다른 요소와 달리 종속성이 높다는 것은 안전도가 낮아지게 되므로 요소 위험지수(FRI)를 1에서 빼준 다음 가중치와 곱해준다.

재현가능성(renderbility) 항목 역시 점수가 높을수록 상대적으로 위험 요소가 작다는 것을 기본으로 한다.

요소 위험지수와 가중치를 기반으로 하여 특정 파일 포맷의 안전도 평가를 위한 안전도 평가 값(safety assessment value: SAV)을 식 (1)과 같이 정의하였다. 안전도 평가 값은 위험 항목 요소에 대한 값을 지수로 변환한 후 가중치를 곱해 합산하여 산정하는 방식이다. 먼저, i번째 전자기록의 안전도 평가 값 SAV_i는 다음의 식 (1)과 같이 구할 수 있다. 또한, 식 (1)의 표준성 평가, 정체성 평가, 종속성 평가 및 재현가능성 평가는 각각 식 (1-1)~식 (1-4)로 표현할 수 있다.

$$SAV_i = f(S_i) + f(I_i) + f(D_i) + f(R_i) \tag{1}$$

$f(S_i)$: *Standardizaion Asse.*, $f(I_i)$: *Identity Asse.*,
$f(D_i)$: *Dependency Asse.*, $f(R_i)$: *Renderability Asse.*

$$f(S_i) = MAX(IS_i, KS_i, CS_i) \tag{1-1}$$

IS_i : *National Standard*, KS_i : *Korean Standard*,
CS_i : *Company Standard*

$$f(I_i) = \sum_{ps \in PS_i} (U_{ps,i} \circ P_{ps,i} \circ A_{ps,i}) \tag{1-2}$$

$$U_{ps,i} = \sum_{j=1}^{n} (U_j * wU_j), \quad P_{ps,i} = \sum_{j=1}^{n} (P_j * wP_j), \quad A_{ps,i} = \sum_{j=1}^{n} (A_j * wA_j)$$

ps : *properity set*
U_j : *understandability FRI*, wU_j : *understandability weight*
P_j : *persistency FRI*, wP_j : *persistency weight*
A_j : *availability FRI*, wA_j : *availability weight*

$$f(D_i) = \sum_{ps \in PS_i} (DM_{ps,i} \circ DS_{ps,i}) \tag{1-3}$$

$$DM_{ps,i} = \sum_{j=1}^{n} ((1 - DM_j) * wDM_j), \qquad DS_{ps,i} = \sum_{j=1}^{n} ((1 - DS_j) * wDS_j)$$

DM_j : *media dependency FRI*, wDM_j : *media dependency weight*
DS_j : *software dependency FRI*, wSM_j : *software dependency weight*

$$f(R_i) = \sum_{ps \in PS_i} R_{ps,i} \tag{1-4}$$

$$R_{ps,i} = \sum_{j=1}^{n} (R_j * wR_j)$$

R_j : *renderability FRI*, wR_j : *renderability weight*

식 (1)에 의해 계산되는 파일 포맷의 안전도 평가 값은 임의의 값으로 표시되므로 이를 파일포맷의 위험여부를 판단하기 위해서는 판단기준이 있어야 한다. 해당파일 포맷의 장기보존을 위한 조치 여부를 결정하기 위해 계산된 안전도 평가 값의 범위에 따라 안전(safety), 보류(hold), 위험(danger)의 3가지 등급(grade)으로 위험도를 표시하도록 한다. 평가 결과 안전 등급인 경우는 현행 파일포맷을 유지하여도 아무런 문제가 없음을 의미하며, 보류 등급은 파일포맷 정보가 소실될 가능성이 존재하다는 것을 뜻한다. 위험 등급은 현행 파일포맷의 장기보존을 위해서는 별도의 조치가 필요함을 의미한다. 등급을 나누는 기준 값으로 LL(LOWER_LIMIT)과 UL(UPPER_LIMIT)의 상수 값을 사용하며 이를 이용한 i번째 파일 포맷의 위험 등급 분류(RG_i)는 식 (2)와 같이 표현 할 수 있다.

$$\begin{aligned} &\text{if } SAV_i \geq UL \text{ then } RG_i = \text{'Safety'} \\ &\text{if } LL \leq SAV_i < UL \text{ then } RG_i = \text{'Hold'} \\ &\text{if } SAV_i < LL \text{ then } RG_i = \text{'Danger'} \end{aligned} \tag{2}$$

3.3 위험 평가 결과 분석

위에서 설명한 방식에 따라 파일 포맷의 위험도를 평가하였다. 파일 포맷에 대한 위험 평가를 수행하는데 있어 위험 평가 요소와 항목 및 분류의 가중치는 대단히 중요한 영향을 미치게 된다. 우리는 DipRec과 AONS 등을 분석 및 참조하여 위험 평가 요소와 항목을 선정하였다. 정량화가 힘든 인기도 등 일부항목은 제외하였으며 국내 현실을 반영하기 위해 국내 표준 등의 항목은 새로 추가하였다. 또한 제시한 21개의 위험 평가 요소를 4개의 카테고리로 그룹화 하였다. 우리의 가중치 결정에는, 유사하게 가중치를 사용하여 평가하는 방식인 DipRec의 가중치를 참조하였다. 그러나 DipRec과 우리의 위험평가 항목이 상이한 부분이 있고 또 평가 방식이 다르므로 DipRec에서 사용한 가중치를 참조하여 전문가 그룹에서 우리의 위험 평가 분류, 항목 및 요소에 맞추어 세분화하고 조정하는 작업을 수행하였다. 여기에서 결정된 위험 평가 요소에 대한 가중치는 속성별 중요도의 차이를 반영할 수 있도록 하였다. 이렇게 적용된 가중치는 Table 2에서 위험 평가 요소 옆에 괄호로 표시하였다. 한편, 등급 분류 기준 값으로 LOWER_LIMIT은 3.5, UPPER_LIMIT은 7.5의 값을 사용하였다. 즉, 0~3.4점 구간은 위험, 3.5~구간은 보류, 7.5점 이상은 안전으로 판단한다. 이 기준 역시 DipRec의 기준 값을 참조하였으나 위험도 변화를 면밀히 추적하기 위해 보류 등급의 비중을 높여 설정하였다.

Table 2에서는 파일 포맷에 대한 위험 평가의 예도 같이 보여주고 있다. 국내에서 많이 사용되는 워드프로세스 파일포맷인 HWP 파일포맷이 어떻게 평가되어지는지를 보여주고 있으며 (식 1)과 (식 2)에 의해 '안전'으로 평가되는 것을 알 수 있다.

본 연구에서 수립한 방식에 대한 신뢰성을 평가하기 위해 선진 사례인 DipRec의 위험평가 알고리즘을 적용하여 평가한 결과 값을 비교해 보았다. DipRec은 파일 포맷에 대한 위험도를 평가하고 있으며 평가 결과는 60점을 만점으로 한다. 0~24 구간을 위험도 Low, 25~35구간은 위험도 Middle 그리고 36~60 구간을 위험도 High로 평가한다[8]. DipRec의 평가 결과는 Fig. 1과 같이 표현되며 MP3는 10점, PDF는 14점, TIF는 26점 등으로 평가됨을 알 수 있다. 또한, mac(구 버전의 맥OS에서 구동한 페인팅 프로그램 포맷), sxw(구 버전의 오픈 오피스 포맷), dxf(구 버전의 Autocad 포맷) 등 현재 사용되지 않는 포맷에 대해서는 위험도 상으로 평가됨을 알 수 있다.

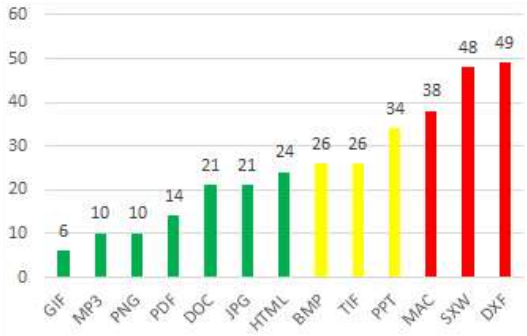

Fig. 1. Risk assessment results by format in DipRec.

본 연구에서 수립한 방식을 이용하여 국내에서 사용 중인 포맷에 대한 안전도 평가를 수행한 결과를 Fig. 2에 나타내었다. 평가 결과는 안전도를 평가하며 최고점은 10점으로 표시된다. DipRec과 달리 위험도가 아니라 안전도를 평가하므로 점수가 낮을수록 위험도가 높게 평가된다. 여기에서 "위험도"란 평가 시점에 해당 파일 포맷의 장기보존과 활용에 위험이 되는 정도를 나타내는 지표가 된다.

평가 결과 국내에서 가장 많이 사용되는 MP3, PDF 등의 포맷은 DipRec에서와 마찬가지로 안전으로 평가됨을 알 수 있다. 한편 국내에서 주로 사용되는 HWP 포맷은 현재 안전으로 평가되나 안전과 보류의 경계선 상에 위치하고 있음을 알 수 있으며, 국내에서 더 이상 사용되지 않는 GUL(훈민정음) 포맷이나 SWF 포맷은 위험으로 평가되고 있음을 알 수 있다.

본 연구에서의 평가 방식은 DipRec의 평가 방식과 다르므로 그 결과를 직접적으로 비교할 수는 없으나

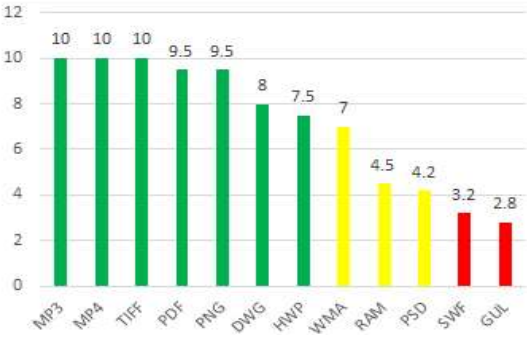

Fig. 2. Risk assessment results in our method.

같은 포맷(MP3, PDF, PNG 등)에 대해서는 같은 단계로 평가됨을 알 수 있으며, DipRec에서는 평가되지 않은 국내 사용 파일 포맷(HWP, GUL 등)의 평가는 이들 포맷의 현재 위험도를 적절히 평가하고 있다고 볼 수 있다.

4. 결 론

본 연구에서 전자기록의 장기보존 의사결정 지원을 위한 위험평가 방식을 개발하여 다양한 포맷을 대상으로 위험 평가를 수행한 결과 다음과 같은 결론 및 향후 연구가 필요하다고 판단된다.

향후 표준 포맷 중심의 기록 보존 체계에 대한 다양한 연구가 필요하다. 또한 현재까지 작성되어 보존 중인 전자기록물의 위험 평가를 위해서는 다양한 위험요소 항목에 대한 기술정보 관리가 필요하며 현재 해외에서 운영 중인 PRONOM에서도 요소 항목에 대한 기술정보가 누락되어 있는 경우가 많으며 특히 국내에서 개발되어 사용된 포맷에 대한 기술정보는 포함되어 있지 않으므로 국가 기관 주도의 포맷 기술정보 관리 체계를 강화할 필요가 있다고 판단되며 이를 기반으로 위험요소 항목의 요소 값을 정의하는 것이 필요하다.

본 연구에서 제시한 21개의 위험요소 항목은 해외 사례를 기반으로 추출된 위험 요소 항목이므로 향후 위험요소 항목에 대한 국내 상황을 반영하는 심도 깊은 연구가 필요하다. 또한 우리가 수행하는 위험 평가라는 작업은 위험이라는 추상적인 요소를 평가값과 등급이라는 정량적 요소로 변환하는 일이므로 평가 기준의 모호성과 평가 결과의 정확성 문제가 따를 수 있으므로 이를 극복하기 위한 노력이 따라야 한다. 그러므로 가중치와 등급 기준 값은 향후 시간의 경과와 기술의 발전에 따라 현실에 맞춰 지속적으로 보정하여 사용하여야 할 것 이다. 가중치와 등급 기준 값의 보정에는 머신러닝 등의 인공지능 기법을 이용한 추가적인 연구를 수행할 계획이다.

지금까지 제시된 위험요소 항목은 현 시점에서 보편화되어 사용되는 파일포맷의 위험도 평가를 위해 필요한 요소이며 IT 기술의 발전으로 ICBM(IoT, Cloud, Bigdata, Mobile) 분야에서 생산되는 다양한 전자기록에 대해서는 지속적인 연구 수행을 통해 위험 요소 항목에 대한 재평가가 주기적으로 이루어져야 한다.

마지막으로 전자기록의 위험요소를 평가하기 위해서는 포맷 기술정보에 대한 평가와 더불어 전자파일 자체에 포함되어 있는 위험요소 즉 압축 여부, 암호화 여부, 렌더링 객체 포함 여부 등을 추가적으로 평가하여야만 개별 전자기록에 대한 위험 평가가 가능하므로 전자파일에 대한 위험평가를 위한 추가적인 연구가 필요하다 하겠다.

REFERENCE

[1] 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.

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

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

[4] W. Sohn, S. Lim, D. Nam and E. Kim, "A study on the Digital Format Registry for digital objects preservation in Korea," *Journal of Korea Multimedia Society,* Vol. 12, No. 10, pp. 1397-1406, 2009.

[5] A. Brown, Head of Digital Preservation, *Automatic Format Identification Using PRONOM and DROID*, The National Archives, 2006.

[6] A. Brown, Services Manager, *Pronom 4 Information Model,* The National Archives, 2005.

[7] D. Pearson, "AONSⅡ : Continuing the Trend Towards Preservation Software Nirvana," *Proceeding of the International Conference on Preservation of Digital Objects 2007*, Beijing China, 2007.

[8] R. Graf and S. Gordea "A Risk Analysis of

File Formats for Preservation Planning," *Proceeding of the 10th International Conference,* pp.177-186, 2013.

[9] S. Gordea, A. Lindley, and R. Graf, "Computing Recommendations for Long Term Data Accessibility Basing on Open Knowledge and Linked Data," *Joint Proceeding of the RecSys*, pp.51-58, 2011.

[10] 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.

[11] S. Vermaaten, "Identifying Threats to Successful Digital Preservation: the SPOT Model for Risk Assessment," *D-Lib Magazine,* 18.(9/10), 2012.

[12] H. Ryan, University of Denver Library and Information Science Program, "Occam's Razor and File Format endangerment Factors," *Proceeding of International Conference on Digital Preservation,* pp. 179-188, 2014.

차 현 철

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

최 주 호

1989년 경북대학교 전자공학과 학사
1991년 경북대학교 컴퓨터공학과 석사
2003년 경북대학교 컴퓨터공학과 박사수료
현재 (주)이산티에스 대표이사