데이터세트 보존포맷 검증방안에 관한 연구: 재난안전정보 데이터세트의 SIARD 적용을 통해*

Empirical Verification of Conversion and Restoration of Preservation Format for Dataset: Application of Dataset with Disaster Safety Information to SIARD

한희정 (Hui-Jeong Han)**, 윤성호 (Sung-Ho Yoon)***
오효정 (Hyo-Jung Oh)****, 양동민 (Dongmin Yang)*****

초 록

정보의 활용이 국가 경쟁력의 핵심으로 부각되면서 우리 정부를 포함한 주요 선진국들은 데이터를 중요하게 인식하고 있으며, 이에 따라 장기보존 기술 연구 및 표준 제정 등을 추진하여 데이터의 체계적인 관리 및 보존을 위한 노력을 지속적으로 기울이고 있다. 그러나 현재 국내의 경우 다양한 유형의 데이터들에 대해 법령에는 기록관리 대상으로 명시하고 있지만, 이를 수집, 관리 및 보존하기 위한 구체적인 방법은 표준전자문서 이외에는 없는 상황이다. 특히, 행정정보시스템에서 생산되는 엄청난 규모의 데이터세트에 대한 관리 및 보존은 무엇보다 강하게 요구되어 왔으나 데이터세트에 대한 지침이 제대로 제공되고 있지 않고 있다. 보존포맷 선정체계가 마련되어야 시스템 보완 및 구축이 가능하기 때문에 우선적으로 데이터세트 특성을 고려한 보존포맷 선정 기준 체계가 보다 구체화 되어야 하며, 선정기준에 따라 도출된 데이터세트 보존포맷의 변환에 대한 실증적인 검증 작업이 필요하다. 이에 본 연구는 데이터세트의 특성을 고려한 보존포맷 선정 기준에 대한 평가체계를 도출하고, 보존포맷에 대한 실증적 검증을 통해 장기보존할 수 있는 방안을 제시하고자 한다.

ABSTRACT

As the use of information has emerged as the core of national competitiveness, major developed countries and the Korean government have realized the importance of data. They have pursued technical research and standard establishment for long-term preservation and continuously strived for systematic management and preservation of data. However, although various types of data are specified for the purpose of record management in the law, there is no specific method on how to collect, manage and preserve them, except standard electronic documents. In particular, management and preservation of huge datasets from the administrative information system have been strongly demanded above all. Any guidelines for datasets do not have been properly provided. After the framework for selecting preservation format must be prepared, the system can be supplemented and built. The framework considering the characteristics of the dataset should be specified more concretely, and empirical verification of the conversion and restoration for the dataset preservation format derived according to the selection criteria is necessary. Therefore, this study intends to propose a method for long-term preservation through empirical verification of the preservation format after deriving an evaluation the framework for the preservation format selection criteria considering the characteristics of the dataset.

키워드: 전자기록 장기보존, 보존포맷 선정체계, 행정정보 데이터세트, 재난안전정보
long-term preservation of electronic records, framework for selection preservation format, administrative information dataset, disaster safety information

---

* 본 연구는 "2019년 행정안전부 국가기록원 기록관리 연구개발사업"의 연구비를 지원받아 수행되었음. 이 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 한국연구재단 - 재난안전플랫폼기술개발사업의 지원을 받아 수행된 연구임(No. NRF-2016M3D7A1912703).

** 전북대학교 문화융복합아카이빙 연구소 전임연구원(freebirdhhj@naver.com) (제1저자)
*** 전북대학교 일반대학원 기록관리학과 석사과정(tjdgh9410@naver.com) (공동저자)
**** 전북대학교 문헌정보학과 부교수, 문화융복합아카이빙 연구소 연구원(ohj@jbnu.ac.kr) (공동저자)
***** 전북대학교 일반대학원 기록관리학과 부교수, 문화융복합아카이빙 연구소 연구원(dmyang@jbnu.ac.kr) (교신저자)

■ 논문접수일자 : 2020년 5월 26일 ■ 최초심사일자 : 2020년 6월 11일 ■ 게재확정일자 : 2020년 6월 22일
■ 정보관리학회지, 37(2), 251-284, 2020. http://dx.doi.org/10.3743/KOSIM.2020.37.2.251

1. 서 론

1.1 연구의 필요성과 목적

최근 데이터의 활용이 국가 경쟁력을 좌우하는 핵심자원으로 그 중요성이 부각됨에 따라 이를 정부차원에서 적극 수집, 관리 및 활용하고자 하는 움직임이 활발해지고 있다. 이미 미국・영국 등 주요 선진국들은 데이터와 통계자료의 중요성을 깨닫고 이들 데이터를 장기보존하기 위한 기술 연구 및 데이터 표준 제정 등을 추진하여 데이터의 체계적인 보존을 위한 노력을 지속적으로 기울여 왔다. 우리나라 역시 2013년에 「공공데이터의 제공 및 이용활성화에 관한 법률」(이하 '공공데이터법')을 제정하여 공공데이터 공유 및 활용 기반을 마련하기 시작했다. 나아가 이를 보다 적극적으로 활성화시키고자 행정안전부는 2017년에 「데이터기반행정 활성화에 관한 법률안」(이하 '데이터기반행정법')을 입법예고한 후 국회에 제출한 바 있다. 데이터기반행정법은 데이터를 기반으로 한 행정의 활성화에 필요한 사항을 규정하여 객관적이고 과학적인 행정을 통하여 공공기관의 책임성, 대응성 및 신뢰성을 높이고 국민의 삶의 질을 향상시키는 것을 목적으로 하는 데이터 관리 전반을 규정하고 있는 법률이다. 여기에서 규정하고 있는 데이터기반행정의 주요 추진 분야는 주요 정책을 수립하거나 경제적・사회적 문제 등을 해결하기 위하여 국민의 의견을 신속하고 정확하게 수렴할 필요가 있는 분야나 안전사고, 질병 등 사전에 위험 요소와 원인을 예측하고 제거방법을 제시할 필요가 있는 분야 등이다(정부, 2017).

공공기관에서 운영 중인 정보시스템에는 국가가 관리하는 인적・물적 자원에 대한 정보, 각종 재난・사고・자연관측 정보, 정보정책과 관련된 행정통계 등 빅데이터 분석에 활용할 수 있는 원천데이터가 포함되어 있다. 이에 행정안전부는 국가적으로 보존가치가 있는 데이터에 대해서는 의무적으로 보존할 수 있도록 전자정부법을 개정하고 공공기관이 실행할 수 있도록 지침을 마련하기 위한 종합계획을 발표한 바 있다(행정안전부, 2018. 9. 19). 이렇듯 데이터 등의 전자기록물을 적극 수집, 관리 및 활용하기 위한 행정안전부의 행보가 이어지면서 기록관리 환경 역시 급속도로 변화함에 따라 국가기록원은 이러한 변화에 적극 대응하기 위해 다양한 방안들을 모색하고 있다. 예컨대, 최근 개정된 「공공기록물 관리에 관한 법률」 제20조 2에 '전자기록물 기술정보의 관리' 조항을 신설하여 전자기록물 장기 보존 및 활용에 필요한 기술정보를 수집할 수 있는 법적 근거를 마련하였다. 뿐만 아니라 동법 시행령 제34조 3에 '행정정보 데이터세트의 관리(이하 데이터세트)' 조항을 신설하여 데이터세트의 관리를 의무화 하였다. 지금까지 공공기관이 보유한 파일포맷 등에 관한 기술정보의 현황파악 및 수집 근거 부족으로 전자기록물 장기보존 전략을 수립하는데 한계가 있었다. 그러나 이번 법개정으로 전자기록물의 장기보존정책을 수립하는데 필요한 전자기록물 기술정보(DFR-Digital Format Registry) 수집 근거 조항이 마련됨에 따라 공공기관이 생산하는 전자기록물의 파일포맷 유형, 재현 및 구동 환경 등 보존 및 활용에 필요한 정보 수집이 가능해지면서 전자기록물을 보다 체계적이고 지속적으로 관리

할 수 있는 길이 열리게 되었다. 특히, 데이터세트를 공공기록물로써 관리할 수 있는 법적 근거가 마련됨에 따라 향후 기록 관리의 범위가 더욱 확장될 가능성이 커졌다.

다만 문제는 현재 데이터세트가 법령으로만 기록관리 대상으로 명시되어 있을 뿐 어떻게 수집, 관리 및 보존해야 하는지에 대한 구체적인 방법은 나와 있지 않다는 점이다. 즉, 현재 철건 구조의 표준전자문서 중심으로 설계되어 있는 기록관리시스템에 데이터세트를 어떻게 관리 범위에 포함시킬 것인지, 장기보존하기 위해 포맷 선정은 어떻게 해야 하는 지 등 데이터세트에 대한 관련 지침이 부재하여 현장에서 많은 혼란과 어려움이 있을 것으로 예상된다. 특히, 공공기관에서 생산되는 많은 다양한 유형의 데이터세트를 지속가능하게 관리 및 보존하기 위해서는 데이터세트 특성을 고려한 보존포맷 선정 기준 체계가 보다 구체화 되어야 하며, 선정기준에 따라 도출된 데이터세트 보존포맷의 변환에 대한 실증적인 검증 작업이 필요하다.

이에 본 연구는 데이터세트의 특성을 고려한 보존포맷 선정 기준에 대한 평가체계를 도출한 후, 보존포맷 변환에 대한 실증적 검증을 통해 데이터세트 유형의 전자기록을 장기보존할 수 있는 방안을 제시하고자 한다.

1.2 연구의 범위와 방법

본 연구는 데이터세트 유형 전자기록의 장기보존 방안을 제안하는 것을 목적으로 한다. 이를 위해 먼저, 문헌조사를 통해 국내외 전자기록물 보존포맷 현황 및 선정 기준을 분석하였으며, 데이터세트 유형 전자기록 분석을 위해 국내 공공기관 행정정보시스템의 형태 및 운영 현황 등을 조사 및 분석하였다. 특히, 공공기관 행정정보시스템 중 최근 사회적 재난의 큰 관심으로 더욱 주목받고 있는 재난・사고・자연관측 등의 재난안전정보를 수집 및 관리하는 RDB형 데이터세트를 중심으로 데이터 유형을 분석하였다. 한 곳으로 수집된 재난안전정보는 빅데이터 및 인공지능 기술 분석을 통해 재난을 예측하고 대응하는데 기초데이터가 되기 때문이다. 먼저, 1~3차 온라인조사에 걸쳐 국민안전처 및 주요 재난안전 유관기관 총 55개 기관의 재난안전정보 현황 및 자동 수집 가능성을 분석하였고, 자동 수집 가능성이 높은 19개 기관(국민안전처, 국가법령정보센터, 국립수산과학원, 산림청, 국립재난안전연구원등) 대상으로 웹페이지에 게시된 정보 중에서 실시간 정보, 계약정보, 입찰, 채용 등을 제외한 정보를 자동 수집하여 MySQL에 총 41개의 테이블을 수집하였다. 다음으로 SP(Significant Properties)를 통해 도출한 데이터세트의 주요 특성을 바탕으로 데이터세트 유형 전자기록 보존포맷 선정을 위한 평가체계를 개발하였다. 그리고 최근 영국, 독일, 덴마크 등 여러 유럽국가에서 참여하고 유럽위원회(EC: European Commision)에서 지원하는 E-ARK 프로젝트, 포르투칼의 RODA 프로젝트, 미국 의회도서관(LOC: Library of Congress) 그리고 국가기록원 등에서 데이터세트 유형 전자기록의 보존포맷으로 SIARD 2.1을 채택 또는 검토하고 있다. 그래서 SIARD 2.1을 대상으로 본 연구에서 개발한 평가체계를 적용하여 보존포맷으로서의 적합성을 검증하였다. 마지막으로 여러 재난안전 관련 공공

기관의 행정정보시스템에서 수집한 RDB형 데이터세트를 SIARD로 변환하고 복원하는 검증 시험을 실시하였다.

1.3 선행연구

그 동안 전자기록의 장기보존에 관한 연구는 지속적으로 진행되어 왔다. 먼저, 국가기록원(2004)은 문서 유형의 전자기록물에 대한 보존포맷 적합성 평가를 실시하여 PDF/A-1을 선정한 바 있다. 그러나 성환혁(2007), 국가기록원(2013)에서는 PDF/A-1이 다양한 유형의 전자기록물의 보존포맷으로는 한계가 있음을 지적하며, 이에 대한 해결 방안이 필요하다고 언급하였다. 이에 따라 다양한 유형의 전자기록에 적합한 보존포맷에 대한 관심이 증가하면서 관련 연구도 꾸준히 진행되었다. 오세라, 정미리, 임진희(2016)는 오피스 유형에 대한 파일포맷으로 XML 기반 개방형 표준인 ODF(Open Document Format)를 고려한 바 있으며, 강현민(2016), 박준영과 이명규(2019), 임나영과 남영준(2019)은 시청각기록물의 이미지를 보존하기 위한 파일포맷 및 디지털화 기준에 대한 연구를 진행하였다.

한편, 그동안 행정정보 데이터세트 기록의 관리 필요성과 시급성에 대해서는 학계에서 지속적으로 언급되어 왔으나 실제 현장에서는 관리 및 보존이 제대로 이루어지지 못했다. 이와 관련하여 현문수(2005)는 기록관리 대상으로 데이터세트를 인식하고 관리할 필요성을 지적하였으며, 이를 위해 영국 TNA의 National Digital Archive of Datasets(NDAD)와 미국 NARA의 Access to Archival Database(AAD)를 비교 분석하여 국가 차원의 데이터세트 관리 및 서비스 사례를 분석한 바 있다. 그러나 점점 종이기록이 아닌 전자기록 중심으로 환경이 변화하고 있고, 전자기록의 유형 또한 급증하는 상황에서 행정정보 데이터세트의 관리 및 보존 방안에 대한 중요성과 시급성이 더욱 강조되고 있다. 이에 따라 최근 행정정보 데이터세트의 관리 및 보존에 대한 연구가 더욱 활발하게 진행되고 있다. 특히, 문서형 기록과 데이터세트 기록을 동일한 방식으로 관리 및 보존하는 것에 대한 문제점을 지적하면서 데이터세트의 기록관리 방안에 대한 연구가 보다 구체적으로 진행되고 있다. 먼저, 왕호성, 설문원(2017)은 전자기록의 단계적 관리와 물리적 보존에 집중하고 있는 현재의 생애주기적 관리체계가 데이터세트 유형의 전자기록에 적용되어서는 안 된다는 점을 강조하면서 데이터세트의 '재현성'에 초점을 두고 데이터세트 기록관리방안을 제안하였다. 또한 오세라, 이해영 (2019) 역시 데이터세트 기록관리가 방치된 가장 큰 원인으로 문서류와 태생이 다른 데이터세트를 문서류 기록과 같은 기준과 관리 방법을 적용하려고 하는 데 있다고 지적하였다. 이에 해당 연구에서는 현장에서 조사한 시스템의 현황 분석 및 실무자 인터뷰를 통해 데이터세트의 관리 기준을 설계하여 현실에서 적용 가능한 관리 절차를 제안하였다. 이들 연구들은 다양한 유형의 전자기록의 관리 및 보존 방안의 필요성에 대해서 강조하고 있으며, 특히 엄청난 속도로 생산되는 행정정보 데이터세트에 대한 기록관리방안을 가장 시급한 문제로 꼽고 있다.

한편, 다양한 유형의 전자기록 보존포맷과 관련하여 송치호, 차현철(2017)과 차현철, 최주

호(2019)는 파일포맷의 위험도를 평가하는 연구를 수행하였으며, 후자의 경우 보존포맷 선정 기준과 평가방식을 체계적으로 제시하였다. 그러나 선정기준 항목들을 선택하게 된 근거가 구체적으로 제시되지 않았으며, 전자기록의 고유한 특성을 반영할 수 없었다. 그리고 한희정, 오효정, 양동민(2020) 연구에서는 전자기록 보존포맷의 선정하기 위하여 모든 전자기록의 유형에 공통적으로 적용할 수 있는 선정기준을 구체적으로 제시하였지만 전자기록의 고유 특성에 대한 평가방식은 다루지 않았다. 데이터세트 보존포맷과 관련하여 소정의(2019)는 필수보존속성(Significant Properties)를 통해 데이터세트의 특성을 도출하고 보존포맷을 선정하기 위한 4개의 기준 항목을 제시하였다는 점에서 의의를 찾을 수 있다. 그러나 SP의 구조(Structure) 특성을 세분화하지 못한 점, 항목들에 대한 명확한 정의가 내려지지 않은 점, 그리고 보존포맷 선정기준 및 평가방식을 제안하지 못한 점을 보완할 필요가 있다.

기존 연구들은 데이터세트의 기록관리 방안을 포괄적으로 제시하는 연구를 수행하였으며, 보존포맷 연구 역시 전자기록의 유형에 상관없이 포괄적으로 적용할 수 있는 방안을 제시하였다. 반면, 본 연구는 선행연구를 확장하여 데이터세트와 같은 특정 유형의 전자기록에 적용할 수 있는 보존포맷 선정을 위한 평가체계를 제안하고, 실제 데이터세트 보존포맷인 SIARD를 대상으로 해당 평가체계를 통해 적합성 및 변환 검증을 실시하여 데이터세트 보존포맷의 검증방안을 보다 실증적으로 제시하였다는 점에서 기존연구들과 차별성을 가진다.

2. 이론적 배경

2.1 데이터세트

데이터세트의 정의를 살펴보면 사람이 아닌 컴퓨터에 의해 처리되는 것을 전제하고 있으며, 다양한 유형(문자, 숫자, 통계, 공간, 서지정보, 이미지 등)의 데이터로 구성되어 있다는 것을 알 수 있다. 그래서 '컴퓨터가 처리하거나 분석할 수 있으며 다양한 형태로 존재하는 관련 정보의 집합체이다.' 정의가 가장 적합하다고 판단된다. 데이터세트는 사람이 아닌 컴퓨터에 의해서만 처리되거나 분석된다는 점이 다른 전자기록물과 구분되는 가장 큰 이유라고 할 수 있다. 컴퓨터가 확인할 수 있으면 되므로 데이터세트의 외관은 전혀 고려되지 않는다. 즉, 문자, 표, 이미지 등의 크기・폰트・색상・음영 등은 중요하지 않고, 표현하고자 하는 내용(문자, 숫자, 기호 등)이 중요하다. 예를 들어, 엑셀이란 응용 프로그램으로 생성된 파일일지라도 데이터만 저장하고 외부의 다른 응용프로그램과 연계되어 사용되는 경우는 데이터세트에 해당되지만, 엑셀을 이용하여 크기・폰트・색상・음영 등을 사용하여 만든 파일의 경우는 데이터세트에 해당되지 않는다. 그래서 데이터세트는 크게 '파일'(JSON, CSV, HTML, SQL, XML, TXT, EXCEL, 한셀, ODS 등) 저장 방식과 "데이터베이스"(Oracle, MySQL, SQL Server, 큐브리드, MongoDB, DynamoDB, DataStax 등) 저장 방식 2가지로 구분될 수 있다. 그리고 파일 저장 방식도 텍스트 파일(Text File) 저장 방식과 문자열 이외에 다른 여러 형태의 데이터를 포함하는 이진 파일(Binary File) 저장 방식

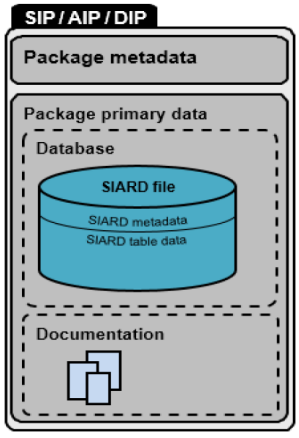
으로 다시 나눌 수 있다. JSON, CSV, HTML, SQL, XML, TXT 등이 대표적인 텍스트 파일 방식이고, EXCEL, 한셀, ODS 등이 대표적인 이진 파일 저장 방식으로 스프레드시트(Spreadsheet)라고 불린다. 데이터베이스도 관계형 데이터베이스 방식과 NoSQL 방식으로 나눌 수 있다. Oralce, MySQL, SQL Server, 큐브리드 등은 관계형이고, MongoDB, DynamoDB, DataStax 등은 NoSQL형이다(노종원, 소정의, 2020).

2.2 SIARD 2.1

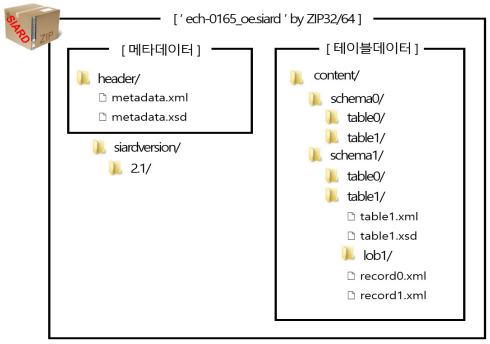
SIARD(Software Independent Archival of Relational Databases)는 관계형 데이터베이스에 저장되어 있는 데이터세트를 소프트웨어와 독립적으로 하나의 파일로 '장기보존' 할 수 있도록 개발된 표준이다. Unicode, XML, SQL:2008, URI(Uniform Resource Identifier), ZIP 등의 표준을 기반으로 하고 있어 원본 데이터베이스 소프트웨어를 사용할 수 없게 되더라도 이들 표준에 기반하여 데이터베이스 데이터에 접근 및 교환이 가능하기 때문에 보존용 포맷으로 고려해 볼 수 있다.

SIARD 개발 현황을 살펴보면 SIARD 1.0은 2007년 SFA(Swiss Federal Archive: 스위스 연방 기록원)에서 개발되어 2013년에 eCH 0165라는 표준으로 제정되었다. 이후 2016년 E-ARK 프로젝트의 일환으로 SIARD 2.0에 이어 현재 2.1까지 나와 있으며 이에 따라 몇 가지 기능이 추가되었다. 예컨대, SQL:2008의 모든 데이터 타입을 지원하며, 사용자 정의 데이터 타입(UDT: User-Defined Data Type)도 사용 가능하게 되었다. 또한 정규표현식(Regular Expression)을 사용하여 데이터 타입 규칙 준수 여부도 검증이 가능해졌다. 그 외에도 데이터베이스 안에 있는 SIARD 파일이 외부에 저장되어 있는 대용량의 객체를 "file:" URI를 이용하여 참조할 수 있으며, 압축방법으로는 deflate 방식을 지원하고 있다.

한편, SIARD는 OAIS 패키지 모델 구조와 독립적으로 설계되어 OAIS 패키지 메타데이터와 관계없이 자체적으로 메타데이터를 가지고 있으며, 다른 문서들(외부 LOB파일, 외부 파일 이름에 대한 변환 맵, DB 문서, DB 구조와 관련 문서 등)과 함께 보존되는 것으로 가정한다(〈그림 1〉 참조). SIARD 아카이브 구조를 보면 메타데이터와 테이블데이터가 결합된 구조로 하나의 관계형 DB는 단일의 SIARD 파일로 저장되며, 모든 DB 콘텐츠는 XML 스키마 1.0의 스키마 정의에 따라 XML 1.0 포맷의 파일 집합으로 보관된다. 스키마 정의와 SQL 코드는 SQL:2008을 따르며 이러한 SIARD 아카이브 내부 파일구조를 표현하면 〈그림 2〉와 같다.



〈그림 1〉 SIARD 정보 패키지
출처: eCH-0165 2018



〈그림 2〉 SIARD 아카이브 내부 파일구조의 개요도(예시)

2.3 필수보존속성: SP (Significant Properites)

필수보존속성은 시간이 경과하여도 보존되어야 하는 디지털 객체의 필수 기능을 특징화한 것으로 디지털 객체가 접근 가능하고 의미 있는 상태를 유지할 수 있도록 시간 경과에 따라 보존되어야 하는 디지털 객체의 중요한 특성이다(Giaretta, Matthews, Bicarregui, Lambert, Guercio, Michetti, & Sawyer, 2009; Knight, 2008). 따라서 필수보존속성을 통해 전자기록에서 보존되어야 할 중요한 특성을 도출하여 보존한다면 기록의 4대 요건을 유지한 상태로 보존할 수 있으며 향후 장기보존 전략을 세우는데 좋은 참고자료로 활용 가능하다(The National Archives, 2018. 5. 1).

디지털 객체에 대한 필수보존속성은 NARA, TNA, PLANETS project, NAA 등 이미 여러 나라에서 연구 및 개발하여 활발히 사용하고 있으며, Essential Characteristics, Significant Characteristics 등과 같이 다른 단어로도 표현된다. 또한 나라마다 필수보존속성의 수와 정의가 차이가 있기도 하다. 예컨대 NARA는 필수보존속성을 Appearance, Behavior, Context, Structure로 구분하였으며(NARA, 2009), TNA는 Rendering, Behavior, Content, Context, Structure로 구분하였다(Knight, 2008). 그리고 NAA는 Appearance, Behavior, Content, Context, Structure로 구분하였다(소정의, 2019 재인용).

필수보존속성을 통해 전자기록에서 보존되어야 할 중요한 특성을 도출하여 보존한다면 기록의 4대 요건을 유지한 상태로 보존할 수 있으며 향후 장기보존 전략을 세우는데 좋은 참고자료로 활용 가능하다(The National Archives, 2018. 5. 1). 이러한 필수보존속성을 데이터세트 관점에서 구분하면 〈표 1〉과 같다(Essen, Rooij, Roberts, & Dobbelsteen, 2011).

〈표 1〉 데이터세트 관점에서 필수보존속성(SP)

범주	데이터세트에서의 의미
Appearance (Rendering)	• 기록 내의 외형적인 모습 - 접근할 수 있는 응용프로그램에서 데이터세트가 화면에 표시되는 방법
Behavior	• 기록의 상호작용 - 접근할 수 있는 응용프로그램에서 상호작용하는 방법
Content	• 기록 내 모든 데이터 및 수식 - 주로 데이터베이스 테이블의 내용이지만 데이터가 화면에 표시되는 방법도 포함될 수 있음
Context	• 기록의 메타데이터 - 데이터베이스를 사용하는 조직, 비즈니스 프로세스에서 데이터를 사용하는 방법 및 응용 프로그램에서 데이터베이스의 정보를 사용하는 방법
Structure	• 기록의 구조정보 및 외부 정보 - 데이터베이스의 데이터: 데이터가 테이블로 구성되고 상호 연결되는 방법

3. 현황분석

3.1 공공기관 행정정보시스템 현황

공공부문의 연도별 정보시스템(누적)은 2015년까지 증가하였지만 그 이후로는 매년 감소 추세인 것으로 나타난다. 정보시스템 도입이 유행처럼 급격하게 증가하면서 대부분 기관에 도입되어 보편화되었기 때문이다(〈표 2〉 참조).

2018년 12월 기준으로 16,531개의 정보시스템은 〈표 3〉처럼 엄청난 규모의 행정정보 데이터세트들을 운영하고 있다. 이러한 행정정보시스템 내 데이터세트는 대부분 DBMS를 통해 수집・저장・관리되며, 그 유형은 정형(숫자, 문

〈표 2〉 정보시스템 연도별 현황

(단위: 개, %)

구분			~2014년	2015년	2016년	2017년	2018년
중앙 행정기관		변동		226	-241	-303	-176
		정보시스템수	2,232	2,458	2,217	1,914	1,738
		증가율		10.13	-9.80	-13.67	-9.20
입사헌법/ 독립기관		변동		0	10	3	-2
		정보시스템수	127	127	137	140	138
		증가율		0	-0.79	-1.46	-0.71
지방 자치단체	광역 자치단체	변동		160	-87	-341	11
		정보시스템수	1,781	1,941	1,854	1,513	1,524
		증가율		8.98	-4.48	-18.39	0.73
	기초 자치단체	변동		-24	-2	-420	-412
		정보시스템수	8,397	8,373	8,371	7,951	7,539
		증가율		-0.29	-0.02	-5.02	-5.18
공공기관		변동		536	-594	-62	-201
		정보시스템수	5,913	6,449	5,855	5,793	5,592
		증가율		9.06	-9.21	-1.06	-3.47
전 체		**변동**		**898**	**-914**	**-1,123**	**-780**
		정보시스템수	**18,450**	**19,348**	**18,434**	**17,311**	**16,531**
		증가율		**4.87**	**-4.72**	**-6.09**	**-4.51**

출처: 행정안전부, 한국정보화진흥원 (2019)

〈표 3〉 '17.7월 행정정보시스템 내 데이터세트 현황조사 결과

구분	산림자원통합관리 시스템	국민신문고 시스템	전자연구노트 시스템	특허넷	국토정보 시스템	화학물질종합정보 시스템
운영기관	산림청	국민권익 위원회	한국과학 기술원	특허청	국토교통부	화학물질 안전원
DB 크기	600M	1.2T	2T	15T	3T(원천데이터) 400G(DW/DM)	329G
테이블수	188개	696개	20개	1,560개	108개	90개

출처: 국가기록원 (2017)

〈표 4〉 DBMS 벤더 현황(2018년 12월 기준)

소프트웨어 유형	벤더명	수량(개)	비율(%)
DBMS	Oracle	12,628	(69.69)
	Microsoft	3,283	(18.12)
	티맥스소프트	907	(5.01)
	큐브리드	947	(5.23)
	AltiBase	355	(1.96)
DBMS 합계		17,603	(100.00)

출처: 행정안전부, 한국정보화진흥원 (2019)

자, 날짜/시간 등)과 비정형(영상, 음성, 이진데이터 등)으로 구분할 수 있다. DBMS 벤더 현황은 〈표 4〉와 같다. 전체의 90% 정도를 Oracle과 SQL Server가 점유하고 있으며, 티맥스소프트(티베로)와 큐브리드(큐브리드), 알티베이스(알티베이스)가 뒤를 잇고 있다. 그러므로 데이터세트 보존포맷의 적합성 검증을 위해서는 DBMS가 저장하고 관리하고 있는 다양한 유형의 데이터 타입들을 완벽하게 보존할 수 있는지를 검증해야 한다.

이 중에서 재난안전정보를 생산, 수집, 관리하고 있는 시스템은 약 277개[1]이다. 또한, 지자체, 중앙부처, 민간에서 각각 별도의 시스템으로 자원을 관리하고 있어 재난이 발생하면 어느 기관에서 어떤 자원을 가지고 있는지 알 수가 없어 자원을 신속하게 동원할 수 없다. 이런 상황을 대비하여 2014년부터 재난관리자원 공동활용시스템 구축사업을 추진하여, 2016년에는 중앙부처와 공사・공단 등 189개 기관으로 확대하였고, 2017년에는 민간단체 19개 팀을 추가하였다. 그리고 2016년부터 KISTI를 중심으로 재난안전정보 공유 플랫폼 기술개발 사업을 통해 재난관리자원 공동활용시스템 연계는 물론 데이터 표준화 및 표준화된 재난안전정보 통합플랫폼을 개발하고 있다.

3.2 SIARD 활용 현황

관계형 데이터베이스 보존을 위해 스위스 연방기록에서 개발한 SIARD는 전 세계 여러 국가에서 사용 또는 연구 중이다(행정안전부, 2018). 이와 관련하여 본 논문에서는 스위스, 덴마크, 포르투칼에서 SIARD를 어떻게 활용하고 있는지에 대해 조사・분석하였다(〈표 5〉 참조).

3.2.1 스위스

스위스의 e-CH 협회(eCH Association)는 스위스 전자 정부의 활성화를 위한 공공-민간 협력 기구로서 SIARD 포맷(eCH-0165)을 제정하였으며, 여기에는 관계형 데이터베이스의 장기보존을 위한 SIARD 파일 형식의 사양이 기술되어 있다. 그리고 이러한 SIARD 포맷 지침을 기반으로 SFA(Swiss Federal Archives: 스위스연방기록원)는 SIARD Toolkot인 'SIARD Suite'를 개발하여 배포하였다. 또한 현재 스위스 주정부 전자 세금데이터의 보관에 대한 모

1) 범정부EA포털 정보화현황에서 현행정보시스템명(재난)과 현행정보시스템운영구분(운영중/개발중/이관)으로 검색한 결과

〈표 5〉 국외 SIARD 활용 현황

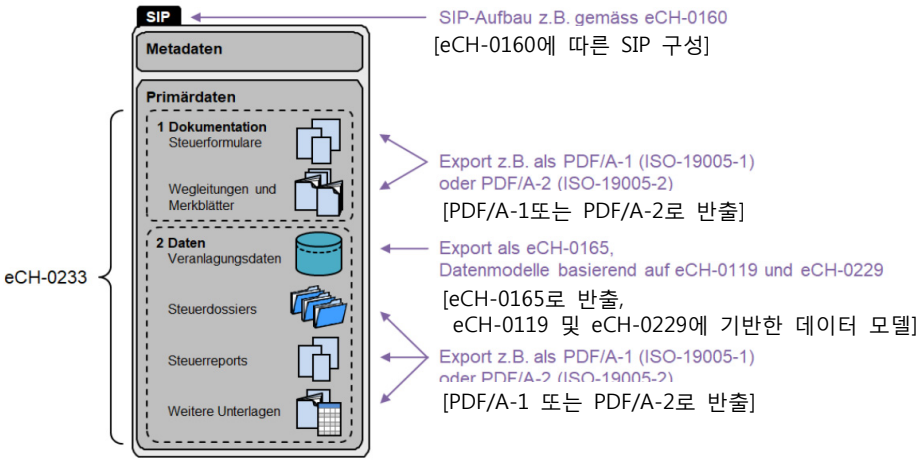
구분	SIARD 현황
스위스	• eCH-0165: SIARD 포맷 지침 • eCH-0233: 주정부 전자 세금데이터의 보관에 대한 모범사례 초안 • SIARD Suite: SIARD Tookit
덴마크	• SIARD-DK: 덴마크 SIARD 표준
포르투칼	• RODA 프로젝트(RODA DBML) • KEEP Solutions • DBPTK

범사례 초안(eCH-0233)이 작성되었으며, 여기에는 입수정보패키(SIP: eCH-0160)로 편집된 세금문서를 다루고 있고, SIARD 파일을 생성하기 위한 데이터 모델 및 SIARD 견본 파일이 첨부되어있어 SIARD 파일을 생성하기 위한 데이터 모델을 확인할 수 있다(〈그림 3〉 참조).

3.2.2 덴마크 국립기록보관소: SIARD-DK (덴마크 SIARD 표준)

덴마크 국립기록보관소(Danish National Archives)는 SIARD 1.0에 기반하여 SIARD-DK를 정보패키지로 사용하고 있다. SIARD-DK는 SIARD 표준 중 하나로 스위스의 SIARD 표준과 유사한 기술 구조를 가지고 있으나, 정보패키지에 대한 덴마크 시행령(bekendtgørelse) 1007/20(2010)이다. 즉, 여기에는 전자기록 입수에 관한 내용이 담겨져 있으며, 특정 멀티미디어 포맷의 확장자 및 저장 상세도 별도로 규정하고 있다(〈표 6〉 참조).

덴마크 기록보관소는 2014-2017년 E-ARK (European Archival Records and Knowledge Preservation) 프로젝트에도 참여하여 SFA과



〈그림 3〉 eCH-0233 개요

출처: eCH-0233 (2019)

〈표 6〉 SIARD-DK 내 명시된 멀티미디어 포맷에 대한 확장자 및 저장상세

멀티미디어 포맷	확장자	저장 상세
TIFF	tif	그래픽 비트맵 TIFF 포맷, version 6.0 baseline (1) 흑백 문서: CCITT / TSS 그룹3, 그룹4, PackBit 또는 LZW로 압축 (2) 그레이 스케일 또는 컬러 문서: PackBit 또는 LZW로 압축
MP3	mp3	DS / EN ISO / IEC 11172-3
MPEG-2	mpg	DS / EN ISO / IEC 13818-2
MPEG-4	mpg	AVC DS / EN ISO / IEC 14496-10 (ITU-T H.264)
JPEG-2000	jp2	ISO / IEC 15444-1: 2004 표준에 따른 JPEG-2000
GML(지리적 특성을 표현하기 위한 XML 포맷)	gml	GML 표준 ISO 19136
WAVE	wav	WAVE LPCM 포맷

함께 SIARD 2 포맷 개발에 참여하였다. 동시에 SIARD 2 포맷으로 데이터를 추출할 수 있는 DBPTK(DataBase Preservation ToolKit) 개발에도 참여하였다.

3.2.3 포르투칼: RODA 프로젝트(RODA DBML), KEEP Solution, DBPTK

포르투칼 국립기록원은 RODA 프로젝트의 일부로 RODA DBML(Database Markup Language)을 개발하였다. RODA DBML은 데이터베이스를 XML Schema인 DBML으로 마이그레이션한 후, 이를 MySQL에 덤프하고 phpMyAdmin으로 시각화하여 접근 권한을 제공하지만, 현재는 새로운 버전이 개발되지 않고 있다. 즉, 새 버전의 DBML은 개발하지 않고, DB의 구조와 컨텐츠를 더 많이 캡쳐하는 SIARD 1.0 개발에 참여하고 있다.

RODA repository는 KEEP Solutions사의 오픈소스 제품이다. KEEP Solutions 사는 포르투갈의 Minho 대학(Universidade do Minho)에서 분사된 디지털 아카이빙 및 디지털 보존 전문회사로서 SIARD 포맷을 위한 DBPTK 또한 KEEP Solutions이 유지보수를 수행하고 있다.

DBPTK(Database Preservation Toolkit)는 데이터베이스를 디지털로 보존하기 위한 데이터베이스 형식 간 변환이 가능한 툴킷이다. 기존의 RODA프로젝트에서 독립되어, E-ARK 프로젝트에서 SIARD 2와 함께 추가로 개발되었다. 특히, SIARD 2로 보존된 경우 Database Visualization Toolkit을 통해 SIARD 파일의 시각화를 지원한다. 그리고 DBVTK(Database Visualization Toolkit)는 SOLR를 기반으로 SIARD 2 파일을 탐색, 검색 및 내보내기가 가능한 툴킷으로 현재 개발 중에 있다.

4. 데이터세트 유형 전자기록물 보존포맷 검증방안

4.1 데이터세트 특성

데이터세트는 컴퓨터가 처리하거나 분석할 수 있는 형태로 존재하는 데이터 자원의 집합체

로써 데이터파일이나 데이터베이스와 종종 동의어로 사용된다(한국기록학회 편, 2008). 데이터세트는 구조화된 정형구조의 데이터베이스뿐만 아니라 첨부한 문서 파일, 동영상, 사진, 음성파일 등과 같은 기능에 의한 다양한 포맷의 파일이 포함되어 있다(오세라, 박승훈, 임진희, 2018). 또한 2019년 기준으로 공공부문에서 보유・운영하고 있는 정보시스템은 총 16,622개이며, 그 중 개별시스템이 총 12,508개, 표준(공통)시스템이 총 3,979개, 단일(공통)시스템이 총135개이다(행정안전부, 2019). 이들 각 시스템마다 다른 데이터 모델, 데이터베이스 종류, 생산하는 데이터의 양을 고려하면 데이터세트는 그 다양성과 복잡성 때문에 관리 및 보존하기가 매우 까다로울 수밖에 없는 기록물이다. 따라서 데이터세트는 그 특성에 따라 관리 및 보존 방법을 다르게 해야 할 필요도 있다. 예컨대, 데이터세트를 보존포맷으로 마이그레이션 할 경우 원래의 기능을 재현하기 위해서는 데이터세트가 DBMS에 복원되어 활용될 수 있느냐는 중요한 문제이다.

이와 관련하여 좀 더 구체적으로 살펴보면 먼저, 데이터세트는 Look&Feel(외형) 보다 데이터의 콘텐츠와 기능이 더 중요하며, SQL문을 통해 외부와의 질의가 이루어지므로 쿼리 또는 외부 링크 또한 매우 중요하게 다루어진다. 그 외에도 데이터 및 수식을 비롯해 문자 인코딩, 템플릿, 스키마 등 사전 정의된 구조 또한 잘 보존되어야 한다. 이러한 점을 고려하여 필수보존속성(SP)을 기준으로 데이터세트의 특성을 도출하면 다음과 같다(〈표 7〉 참조).

먼저, 필수보존속성인 'Structure' 관점에서 살펴보면, 데이터세트는 관계성(Relationship)을 특징으로 한다. 예컨대, 데이터세트 중 관계형 데이터베이스는 기본적으로 테이블로 구성되어 있으며, 여러 테이블들은 하나의 스키마 또는 데이터베이스에 포함되는 등 테이블 간에는 관계(Relationship: PK/FK)가 존재한

〈표 7〉 데이터세트 특성

SP	특성 설명	데이터세트 특성
Structure	• 관계형 데이터베이스는 기본적으로 Table(Column, Row)로 구성됨 • 테이블 간 관계(Relationship: PK/FK)이 존재하며, 여러 Table은 하나의 Schema 또는 Database에 포함되기도 함	관계성 (Relationship)
	• 대부분 관계형 데이터베이스는 이러한 구조를 가지고 있으며, 이 구조는 반드시 보존되어야 할 필수보존 속성임 • 상용화된 데이터베이스들은 이러한 구조를 각자 다른 설계를 통해 구현하고 있으며, 데이터베이스는 지속적으로 업데이트되기 때문에 여러 버전들이 존재함	다양성 (Diversity)
Content	• 데이터베이스의 규모가 클수록 기능이 다양해지므로 관련 데이터세트 요소가 증가하고 복잡해짐. 이러한 데이터 뿐 아니라 프로시저 등과 같은 루틴(Routine)도 필수보존 속성이며 Content 특성에 대응됨	복잡성 (Complexity)
	• 데이터세트는 정형 데이터뿐만 아니라 전자문서 및 이미지 파일과 같은 비정형 데이터, 여러 가지 데이터타입이 데이터세트 내에 포함되므로 필수보존 속성이며 Content 특성에 대응됨	이질성 (Heterogeneity)
Behavior	• 데이터세트는 생산 후, 계속해서 활용되며 SQL문을 통하여 데이터를 이용하기 쉽도록 선별 및 조합될 수 있으므로 필수보존 속성의 Behavior 특성에 대응됨	상호작용성 (Interactivity)

다. 예를 들어, Schema, Table, Column, Row, Relationship(PK/FK) 등이 이에 해당한다. 그 외에도 구조적 관점에서 데이터세트는 다양성을 특징으로 한다. 예컨대, 현재 상용화된 데이터베이스들은 독자적으로 구현되어 지속적으로 업데이트가 되기 때문에 각 데이터베이스들마다 여러 버전들(Oracle(v5, v6, ..., 10g, 11g, 12c, ...), MySQL(1, 2, ... , 8, ...), SQL Server(2013, 2017, 2019, ...), Maria DB(5.1.x, 5.2.x, ..., 10.5.x, ...), CUBRID(1, 2, ..., 11, ...), 등)이 다양하게 존재하게 된다. 따라서 필수보존속성인 구조적인 관점에서 데이터세트의 특성을 종합하면 관계성과 다양성을 도출할 수 있다.

둘째, 필수보존속성인 'Content' 관점에서 보면 데이터세트는 이질성과 복잡성을 특징으로 한다. 이는 데이터세트 안에 여러 이질적인 데이터타입이 포함되어 있다. 또한, 규모나 기능이 다양해질수록 데이터세트 관련 요소도 증가하고 복잡해지기 때문이다. Privilege, User, Stored Procedure, Function, Partitions, Role, Trigger, View, Index 등이 대표적인 예이다. 또한, 데이터세트에는 정형 데이터뿐만 아니라 전자문서 및 이미지 파일과 같은 비정형데이터 등 여러 가지 데이터타입(정수형(INT, SHORT), 실수형(FLOAT, DOUBLE), 문자형(CHAR, VARCHAR), 문장형(STRING, CLOB), 바이너리형(BLOB), 시간형(DATE, TIME) 등)이 데이터세트 내에 포함되어 있다.

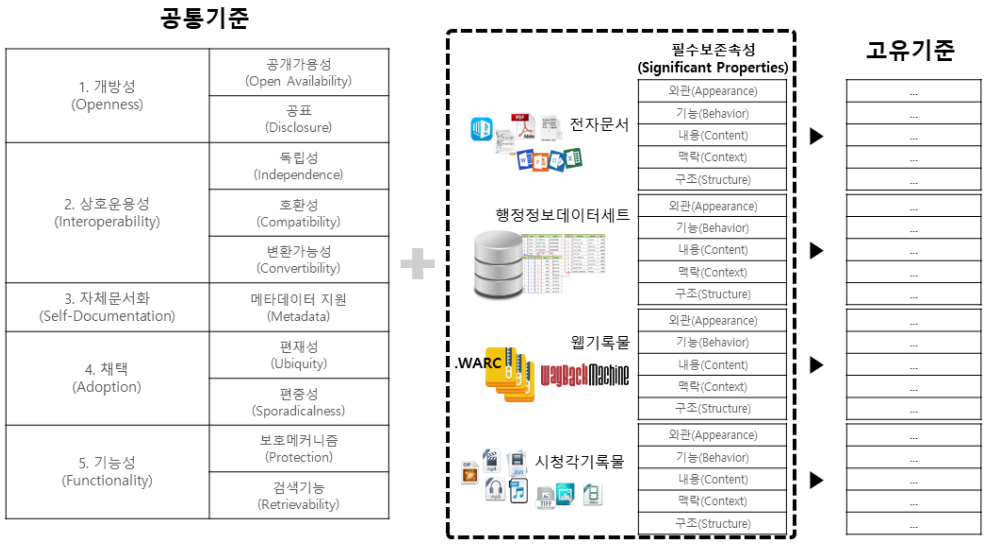
셋째, 필수보존속성인 'Behavior' 관점에서 데이터세트는 상호작용성을 특징으로 한다. 즉, 데이터세트는 생산 후에도 계속해서 활용될 뿐만 아니라 SQL문을 통해 데이터의 선별 및 조합을 통해 데이터를 이용할 수 있다. SQL문은 SELECT, JOIN, CREATE, INSERT, UPDATE, ALTER, GRANT, DELETE, DROP, SHOW 등의 다수의 명령어로 이루어져 있다.

한편, 본 논문에서는 데이터세트 특성을 도출하는데 필수보존속성인 'Appearance'와 'Context'는 고려하지 않았다. 'Appearance'의 경우 전자문서와는 달리 데이터세트에서는 폰트와 레이아웃 등 외형적 요소보다는 보다는 데이터 및 기능의 보존가치가 더 높기 때문이다. 그리고 메타데이터와 관련된 'Context'의 경우 전자문서 등 다른 전자기록과 구별되는 데이터세트만의 고유 특성을 도출할만한 부분이 없었기 때문이다.

4.2 데이터세트 보존포맷 평가체계

4.2.1 전자기록물 보존포맷 평가 체계

전자기록물에 대한 지속가능한 접근을 보장하기 위해서는 적절한 장기보존포맷으로 변환하여 관리해야 한다. 이를 위해서는 먼저 전자기록물 유형별 보존포맷을 선정하는 체계가 마련되어야 한다. 즉, 장기보존포맷으로서 모든 전자기록물 유형에 공통적으로 적용되어야 할 공통기준과 전자기록물 유형별로 보존될 특성에 따라 고려되어야 할 고유기준이 모두 필요하다. 이에 대해 좀 더 구체적으로 살펴보면 먼저, 공통기준의 경우 전자기록물이 담겨진 파일이 SW와 HW에 의해 생성, 저장, 구동되는 기본 원리는 유사하다고 보고, 기록 유형에 상관없이 공통적으로 장기보존에 적합한 기술로서 선정될 수 있는 기준이다. 이와 관련하여 한희정, 오효정, 양동민(2020)은 장기보존포맷으



〈그림 4〉 전자기록물 선정기준

출처: 한희정, 오효정, 양동민 (2020)

로 고려되어야 할 공통기준을 총 5가지를 제시한 바 있다.

그러나 공통기준은 IT 관점에서 장기보존에 적합한 기술에 중점을 둔 것으로 기록관리적인 관점에서 전자기록물 유형별로 어떤 특성을 보존해야 하는지에 대해서는 고려되지 않았다. 따라서 기록관리적 측면에서 공통기준 외에 전자기록물 유형별 보존되어야 할 특성을 분석하고, 이를 반영한 고유기준이 마련되어야 한다.

이에 본 연구는 기록관리적 측면에서 전자기록물 유형 중 데이터세트에 대한 특성을 분석하여 데이터세트 보존포맷 선정을 위한 고유기준을 다음과 같이 제안하고자 한다.

4.2.2 데이터세트 보존포맷 선정을 위한 고유기준

앞 장에서 언급된 필수보존속성(SP)은 진본성 보장을 위해 고려되어야 할 전자기록의 특성을 도출하는데 활용되는 틀이다. 따라서 필수보존속성(SP)을 기준으로 도출한 데이터세트의 특성은 진본성을 보장을 위해 고려해야 할 중요한 요소라 볼 수 있다.

이에 본 논문은 필수보존속성(SP)을 기준으로 도출한 데이터세트의 특성을 바탕으로 데이터세트 보존포맷으로 선정하기 위해 고려해야 할 기준 항목을 다음과 같이 제안하고자 한다(〈표 8〉 참조).

먼저, 데이터세트 보존포맷 선정을 위해 고려해야 할 첫 번째 고유기준은 일반화(Normalization)이다. 일반화는 상용화된 다양한 종류(제조사, 버전)의 DBMS와의 호환가능성을 판단하는 기준이다. 앞서 언급하였듯이 데이터세트는 매우 다양하다는 특성을 갖고 있다. 따라서 데이터세트의 다양성을 고려하여 보존포맷을 선정할 필요가 있다. 특히, 보존포맷이 오픈소스 일

〈표 8〉 데이터세트 보존포맷 선정을 위한 고유기준 항목 및 설명

데이터세트 특성	고유기준	내용	
다양성 (Diversity)	일반화 (Normalization)	정의	• 보존포맷은 상용화된 다양한 종류(제조사, 버전)의 DBMS와 호환이 가능해야 한다는 기준
		설명	• 보존포맷이 오픈소스일 경우, 지원하지 않은 DBMS를 호환 가능하게 하는 것이 중요
관계성 (Relationship) 복잡성 (Complexity) 이질성 (Heterogeneity)	수용성 (Acceptability)	정의	• DBMS의 다양한 현재 그리고 미래에 추가될 데이터 구조 및 관계, 데이터 타입(정형/비정형) 및 루틴 타입을 수용할 수 있어야 한다는 기준
		설명	• DBMS 데이터세트 내 테이블 구조 및 관계, 데이터 타입 (문자/숫자/문장/이진형 등), 루틴 타입(Stored Procedure, Function, Trigger), External File(비정형 데이터) 등을 수용 및 보존해야 함
상호작용성 (Interactivity)	활용성 (Usability)	정의	• 데이터세트를 보존포맷으로 변환 후 활용 가능해야 한다는 기준
		설명	• 보존포맷을 다시 DBMS로 복원하지 않고, 보존포맷 그대로 활용할 수도 있어야 함(예, 뷰어) • 데이터세트가 보존포맷에서 재현을 위해 DBMS 복원가능 해야 함

경우에는 지원하지 않는 DBMS와 호환이 가능하게 하는 것이 무엇보다 중요하다.

다음으로 데이터세트 보존포맷 선정을 위해 고려해야 할 두 번째 고유기준은 수용성(Acceptability)[2]이다. 수용성은 DBMS의 다양한 현재 그리고 미래에 추가될 데이터 타입(정형/비정형) 및 루틴 타입의 수용가능성을 판단하는 기준이다. 데이터세트의 구조는 기본적으로 테이블(column, row)로 구성되어 있어 테이블 간 관계가 존재하며, 데이터베이스의 규모가 클수록 기능과 요소가 증가하고 복잡해지는 특성을 갖는다. 또한 정형데이터 뿐만 아니라 비정형데이터 등 여러 가지 이질적 데이터타입이 데이터세트 내에 포함되어 있어 이를 고려한 보존포맷이 선택되어야 한다. 따라서 DBMS 데이터세트 내 데이터 타입(문자, 숫자, 문장, 이진형 등), 루틴 타입(Stored Procedure, Function, Trigger 등), External File(비정형 데이터)등을 수용 및 보존할 수 있는 보존포맷이 선택되어야 한다.

마지막으로 데이터세트 보존포맷 선정을 위해 고려해야 할 세 번째 고유기준은 활용성이다. 활용성은 변환된 데이터세트 보존포맷의 활용가능성을 판단하는 기준이다. 데이터세트는 생산 후에도 지속적으로 활용되며, SQL문을 통해 데이터를 선별 및 조합하여 이용의 편의성과 용이성을 높일 수 있기 때문에 상호작용적 성격이 강하다. 따라서 이러한 상호작용성을 고려하여 보존포맷이 선택되어야 한다. 예컨대, 데이터세트의 보존포맷은 다시 DBMS로 복원하지 않고도 뷰어와 같이 보존포맷 그대로 활용할 수 있거나 데이터세트가 보존포맷에서 재현될 수 있도록 DBMS로의 복원이 가능해야 한다.

2) 수용성(Acceptability)의 경우 아직 용어에 대한 검증이 이루어지지 않았으며, 다른 후보 용어로서 'Convertibility', 'Transferability' 등이 있음.

4.2.3 데이터세트 보존포맷 평가항목

앞 장에서는 필수보존속성(SP)을 기준으로 도출한 데이터세트의 특성을 바탕으로 데이터세트 보존포맷을 선정하는데 고려해야 할 고유기준 총 3개를 도출하였다. 이렇게 도출된 고유기준은 데이터세트 보존포맷 선정기준의 평가체계 구축을 위한 평가항목으로서 고려할 수 있다. 예컨대, 데이터보존 포맷이 일반성을 확보하였는지를 판단하기 위해 해당 DBMS와 보존포맷의 변환가능성과 변환 소프트웨어가 오픈소스로 존재하는지를 평가할 수 있어야 한다.

이와 관련하여 본 논문에서 제안한 평가항목 중 일반성의 (1-3)번 항목은 2018/2019년도 정보자원 현황 통계 보고서의 통계에 따라 국내 DBMS 소프트웨어 상위 5개(국내 시장의 100%)의 변환가능성을 평가 기준으로 보았으며, (4) 실제성을 확인하기 위해 오픈소스의 제공여부 검증도 필요하다.

수용성과 관련하여 데이터세트는 이질적이고 복잡한 데이터를 수용할 수 있어야 한다. 따라서 RDB형 데이터베이스가 제공하는 내용을 데이터세트 보존포맷이 (5) 테이블 구조 및 관계, (6) 다양한 데이터타입(문자・숫자・날짜・시간 등의 기본 데이터타입과 집합・리스트・목록 등 특수 데이터타입), (7) 루틴계열 타입(Trigger, Function, Stored Procedure 등), (8) 데이터베이스와 연결되어 내부 또는 외부 서버의 디스크에 별도로 저장되어 있는 external file을 보존할 수 있는지를 평가하는 것은 중요하다.

RDB형 데이터세트의 가장 중요한 특징은 외부의 다른 요소들과 SQL로 연계되어 활용되는 것이 핵심기능으로 DBMS에 탑재되어 있을 때에만 그 역할을 수행할 수 있다. 그러므로 RDB형 데이터세트 보존포맷은 기본적으로 데이터세트를 확인할 수 있는 것은 물론, 보존포맷 자체에서 SQL 기능을 수행할 수 있거나 SQL을 수행할 수 잇도록 DBMS에 복원될 수 있도록 설계되어야 한다. 그래서 활용성과 관련하여 데이터세트를 보존포맷으로 변환한 후에도 다시 복원하여 활용할 수 있거나 뷰어와 같은 도구를 통해 확인할 수 있는지를 판단할 수 있는 평가항목(9-15)이 필요하다.

이를 종합하여 본 논문에서는 데이터세트 보존포맷 선정을 위한 평가체계를 〈표 9〉와 같이 제안하고자 한다. 본 논문에서 제안하는 데이터세트 보존포맷 평가표는 다른 전자기록에도 공통으로 적용될 수 있는 공통기준을 제외한 데이터세트만의 특성을 반영한 결과이다.

4.3 데이터세트 보존포맷 검증

앞서 제안한 데이터세트 보존포맷 평가체계를 검증하기 위해 최근 RDB형 데이터세트 보존포맷으로서 거론되고 있는 SIARD를 대상으로 4.3.1에서는 적합성 검증을 수행하였다. 또한, 적합성 검증 결과를 실험적으로 확인하기 위해서 4.3.2에서는 SFA에서 제공하는 SIARD Suite 오픈소스를 활용하여 실제 데이터세트의 변환 및 복원 검증을 실험적으로 실시하였다.

4.3.1 SIARD 적합성 검증

SIARD를 대상으로 데이터세트 보존포맷으로서의 적합성을 검증하기 위해 본 논문 〈표 9〉에서 제시한 데이터세트 보존포맷 선정을 위한 고유기준(평가 결과는 〈표 11〉 참조)과 한희

정, 오효정, 양동민(2020) 논문에서 제시한 공통기준과 평가방식(평가 결과는 〈표 10〉, 등급과 평점 기준은 〈표 12〉 참조)을 보존포맷 선정체계 적용하였으며, 그 결과는 다음과 같다.

〈표 9〉 데이터세트 보존포맷 평가표

고유기준	평가 항목		Y/N
일반화 (Normalization)	1	5개 이상의 DBMS의 데이터세트를 해당 보존포맷으로 변환 가능한가?	Y/N
	2	3개 이상의 DBMS의 데이터세트를 해당 보존포맷으로 변환 가능한가?	Y/N
	3	1개 이상의 DBMS의 데이터세트를 해당 보존포맷으로 변환 가능한가?	Y/N
	4	보존포맷 변환 SW가 오픈소스로 존재하는가?	Y/N
수용성 (Acceptability)	5	보존포맷은 데이터세트의 테이블구조(Column, Row) 및 관계(Relationship)를 보존할 수 있는가?	Y/N
	6	보존포맷은 데이터세트의 데이터타입 계열(문자형, 숫자형, 날짜형, 이진형, 대용량 등)의 데이터를 보존할 수 있는가?	Y/N
	7	보존포맷은 데이터세트의 루틴타입 계열(Stored Procedure, Function, Trigger 등)을 보존할 수 있는가?	Y/N
	8	보존포맷은 데이터세트의 External File을 보존할 수 있는가?	Y/N
활용성 (Usability)	9	보존포맷은 데이터세트의 활용을 위하여 뷰어와 같은 도구를 통해 데이터세트를 확인할 수 있는가?	Y/N
	10	보존포맷은 데이터세트의 활용을 위하여 뷰어와 같은 도구를 통해 SQL 수행이 가능한가?	Y/N
	11	보존포맷은 원래의 DBMS[3]로 테이블구조(Column, Row) 및 관계(Relationship)를 복원할 수 있는가?	Y/N
	12	보존포맷은 원래 생성된 DBMS로 데이터타입 계열(문자형, 숫자형, 날짜형, 이진형, 대용량)을 복원할 수 있는가?	Y/N
	13	보존포맷은 원래 생성된 DBMS로 루틴타입 계열(Stored Procedure, Function, Trigger 등)을 복원할 수 있는가?	Y/N
	14	보존포맷은 원래 생성된 DBMS로 External File을 복원할 수 있는가?	Y/N
	15	보존포맷은 원래 생성된 DBMS가 아닌 다른 DBMS로 복원할 수 있는가?	Y/N

〈표 10〉 전자기록 보존포맷으로서 공통기준 적합성 평가: SIARD

공통기준		평가항목		Y/N	점수
개방성	1. 공개가용성	1.1 특정 기업 외 해당 포맷을 구동시킬 수 있는 다른 SW가 있는가?		Y	1
		1.2 해당 포맷 사용에 대한 제한 여부(라이센스, 구독, 특허료 등)	1.2.1 무료 Read인가?	Y	1
			1.2.2 무료 Write인가?	Y	1
		1.3 기본 도구(메모장, 그림판 등) 사용을 통한 분석가능 여부	1.3.1 기본 도구를 통해 해당 포맷을 구성하는 콘텐츠 전체를 해석할 수 있는가?	Y	1
			1.3.2 텍스트 콘텐츠가 표준 문자 인코딩(UTF-8, 유니코드, 아스키 코드 등)으로 되어 있는가?	Y	1
			1.3.3 압축되어 있는 경우 신뢰성 있는 압축(zip, gzip, lwz 등)으로 되어 있는가?	Y	1
			1.3.4 멀티미디어 콘텐츠가 공개 포맷(jpeg, gif, mpeg 등)으로 되어 있는가?	N	0

3) 원래의 DBMS는 최초로 생성되었던 데이터세트가 관리되었던 DBMS와 동일한 기종이며, 해당 기종의 버전을 지원하는 DBMS를 의미함.

공통기준		평가항목		Y/N	점수
개방성	2. 공표	2.1 해당 포맷의 '표준' 존재 여부	2.1.1 해당 포맷의 표준을 인터넷 등을 통해 공개적으로 참조 및 이용이 가능한가?	Y	1
			2.1.2 해당 포맷의 표준을 인터넷 등을 통해 공개적으로 참조 및 이용할 때 무료인가?	Y	1
			2.1.3 체계적이고 권위 있는 기관에 의해 표준화 과정을 거쳤는가?	Y	1
		2.2 해당 포맷의 '공개코드' 존재 여부	2.2.1 해당 포맷이 오픈소스 라이선스인가?	Y	1
상호 운용성	3. 독립성	3.1 OS 관점	3.1.1 해당 포맷을 구동할 수 있는 OS의 개수가 다수인가?	Y	1
		3.2 HW 관점	3.2.1 해당 포맷을 특별한 HW없이 구동할 수 있는가?	Y	1
			3.2.2 해당 포맷을 개인용 컴퓨터 수준의 HW에서 구동할 수 있는가?	Y	1
		3.3 특정 기술, 표준, 부가SW	3.3.1 해당 포맷 또는 구동 SW에 특수 코덱 및 특수 플레이어와 같은 특정 기술이나 부가 SW 등의 영향이 없는가?	N	0
	4. 호환성	4.1 해당 포맷이 현재 구동 SW에서 지원하는가? (동일한 SW(같은 제조사, 계열사, 인수회사 등)에 한함)		Y	1
		4.2 해당 포맷이 이전/이후 구동 SW 버전과 호환이 가능한가? (동일한 SW(같은 제조사, 계열사, 인수회사 등)에 한함)		Y	1
		4.3 해당 포맷은 구동하는 SW의 Release 주기(공개 주기)에 따라 형식이나 사양이 자주 업데이트되는가? (현재 가장 대표성 있는 구동 SW)		Y	1
		4.4 해당 포맷의 버전 업데이트 개발 로드맵 또는 계획이 존재하는가?		N	0
	5. 변환가능성	5.1 보존, 추후 안정적인 마이그레이션 보장 가능성	5.1.1 해당 포맷이 정보의 손실없이 다른 포맷으로 변환 가능한가?	Y	1
			5.1.2 변환 가능한 포맷이 다양한가?		
		5.2 해당 포맷을 활용하기 쉬운 포맷으로 변환가능 여부 (AIP → DIP)	5.2.1 해당 포맷이 SW, 서비스 및 툴과 상호 운용되어 새로운 목적으로 콘텐츠를 조작하고 재사용할 수 있는가?	N	0
자체 문서화	6. 메타데이터 지원	6.1 해당 포맷이 자동 생성 메타데이터 기능을 제공하는가?		Y	1
		6.2 해당 포맷이 사용자 지정 메타데이터 기능을 제공하는가?		Y	1
		6.3 해당 포맷으로부터 메타데이터를 추출할 수 있는 기능을 지원하는가?		Y	1
채택	7. 편재성	7.1 OS에서 별도의 응용 SW 설치 없이 해당 포맷을 인식하고 내용을 확인할 수 있는가?		N	0
		7.2 브라우저 (Microsoft Edge, Internet Explorer, Chrome, Firefox 등)에서 별도의 확장 응용 SW 설치 없이 해당 포맷을 인식하고 내용을 확인할 수 있는가?		N	0
		7.3 해당 포맷이 표준화 단체에 의해 표준화 과정을 거쳐 저명한 컨소시엄과 그룹에 의해 채택되어 전 세계에서 사용하는가?		N	0
		7.4 해당 포맷이 시장을 선도하는가?		Y	1
		7.5 해당 포맷을 제작/조작/렌더링하는 많은 경쟁 제품의 존재하는가?		N	0
기능성	8. 보호메커니즘	8.1 해당 포맷이 암호 보호, 복사 방지, 디지털 서명, 인쇄 방지 및 컨텐츠 추출 보호와 같은 기술보호메커니즘이 적용되어 있지 않은가?		N	0
		8.2 해당 포맷이 오류 감지, 수정 메커니즘 및 암호화 옵션을 수용하는가?		N	0
		8.3 해당 포맷이 우발적인 손상에 대한 탄력성이 있는가?		N	0
	9. 검색기능	9.1 해당 포맷이 이용자가 원하는 문서내용에 대한 검색 기능을 제공하는가?		Y	1
합계					22/33

〈표 11〉 RDB형 데이터세트 보존포맷으로서 고유기준 적합성 평가: SIARD

고유기준		평가 항목	Y/N	점수
일반화	1	5개 이상의 DBMS의 데이터세트를 해당 보존포맷으로 변환 가능한가?	Y	1
	2	3개 이상의 DBMS의 데이터세트를 해당 보존포맷으로 변환 가능한가?	Y	1
	3	1개 이상의 DBMS의 데이터세트를 해당 보존포맷으로 변환 가능한가?	Y	1
	4	보존포맷 변환 SW가 오픈소스로 존재하는가?	Y	1
수용성	5	보존포맷은 데이터세트의 테이블구조(Column, Row) 및 관계(Relationship)를 보존할 수 있는가?	Y	0.5[4]
	6	보존포맷은 데이터세트의 데이터타입 계열(문자형, 숫자형, 날짜형, 이진형, 대용량 등)의 데이터를 보존할 수 있는가?	Y	1
	7	보존포맷은 데이터세트의 루틴타입 계열(Stored Procedure, Function, Trigger 등)을 보존할 수 있는가?	N	0
	8	보존포맷은 데이터세트의 External File을 보존할 수 있는가?	Y	1
활용성	9	보존포맷은 데이터세트의 활용을 위하여 뷰어와 같은 도구를 통해 데이터세트를 확인할 수 있는가?	Y	1
	10	보존포맷은 데이터세트의 활용을 위하여 뷰어와 같은 도구를 통해 SQL 수행이 가능한가?	Y	0.5[5]
	11	보존포맷은 원래의 DBMS로 테이블구조(Column, Row) 및 관계(Relationship)를 복원할 수 있는가?	Y	1
	12	보존포맷은 원래 생성된 DBMS로 데이터타입 계열(문자형, 숫자형, 날짜형, 이진형, 대용량)을 복원할 수 있는가?	Y	1
	13	보존포맷은 원래 생성된 DBMS로 루틴타입 계열(Stored Procedure, Function, Trigger 등)을 복원할 수 있는가?	N	0
	14	보존포맷은 원래 생성된 DBMS로 External File을 복원할 수 있는가?	Y	1
	15	보존포맷은 원래 생성된 DBMS가 아닌 다른 DBMS로 복원할 수 있는가?	Y	1
합계				12/15

〈표 12〉 전자기록 보존포맷 등급 및 평점 기준

등급	평점(환산점수)	수준정의
A (매우 우수)	90 이상	• 매우 높은 수준의 안정적인 전자기록 보존포맷 • 보존포맷 적합성: 적합 - 10년마다 재평가 실시하여 등급 재설정
B (우수)	80 이상 (80이상 ~ 90미만)	• 높은 수준의 전자기록 보존포맷이지만 정기적인 평가 필요 • 보존포맷 적합성: 적합 - 5년마다 재평가 실시하여 등급 재설정
C (양호)	70 이상 (70이상 ~ 80미만)	• 전자기록 보존포맷으로 선정하기에는 다소 미흡한 부분이 있으므로 보존포맷 선정 여부는 상대평가로 결정 • 보존포맷 적합성: 부분적합 - B 등급 이상의 보존포맷이 없거나 적은 경우 채택 - 3년마다 재평가 실시하여 등급 재설정
D (보통)	60 이상 (60이상 ~ 70미만)	• 전자기록 보존포맷으로 선정하기에는 상당히 미흡한 부분이 있으므로 보존포맷 선정 여부는 상대평가로 결정 • 보존포맷 적합성: 부분적합 - C 등급 이상의 다른 보존포맷이 없는 경우에만 채택 - 3년마다 재평가 실시하여 등급 재설정
E (미흡)	60 미만	• 전자기록 보존포맷으로서 매우 미흡하므로 선정 불가 • 보존포맷 적합성: 부적합

출처: 한희정, 오효정, 양동민(2020)

4) Oracle에서는 관계 보존이 안되므로 부분점수 부여.
5) 별도의 XML뷰어 개발을 통해 일부 기능 구현 가능.

위의 결과를 종합하면, RDB형 데이터세트 보존포맷으로서 SIARD를 선정하기에는 다소 부족한 부분이 있는 것으로 판단되었다. 예컨대, 수용성 기준과 관련하여 일부 데이터베이스에서 관계 보존이 안되었으며, 루틴타입 계열은 보존할 수 없는 것으로 확인되었다. 그리고 활용성 부분에서 SIARD는 뷰어와 같은 도구를 통해 SQL문 실행이 불가능하며, 원래 생성된 DBMS로 External File 복원이 어려웠다. 그 외에도 공통기준에서 부족한 부분이 일부 발견되었다. 따라서 SIARD를 전자기록 보존포맷으로서 공통기준과 데이터세트 특성을 반영한 고유기준을 정성적으로 평가한 결과 SIARD는 관계형 데이터세트 보존포맷으로서 '부분적합' 결과가 도출되었다(〈표 13〉 참조). 그러나 전자문서와는 달리 현재 RDB형 데이터세트의 보존포맷은 많지 않은 편이다. 따라서 SIARD를 RDB형 데이터세트 보존포맷으로 고려하되 정기적인 모니터링을 통해서 지속적인 검증을 시행할 필요가 있다.

본 논문은 RDB형 데이터세트 보존포맷으로서 SIARD의 적합성 평가를 정성적으로 진행하였으며, 가중치는 고려하지 않았다. 따라서 향후 평가의 객관성과 정확성을 높이기 위해 각 평가항목마다 가중치를 부여하여 보다 정밀하게 정량평가가 진행될 필요가 있다.

4.3.2 SIARD 변환 및 복원 검증

보존의 목표는 생산 당시의 모습과 기능을 있는 그대로 재현하는 것으로 RDB형 데이터세트의 경우에는 데이터세트를 DBMS에 복원이 가능해야 본래의 모습과 기능을 재현할 수 있다. 또한, 적합성 평가와 관련해서는 다수의 DBMS를 지원 가능한지(일반화), 다양한 데이터 타입을 보존할 수 있는지(수용성), 기능 재현을 위해 해당 DBMS에 복원가능한 지(활용성)에 대해서 실험적 검증이 필요하다. SIARD 변환 및 복원 검증은 DBMS에서 SIARD 포맷으로 변환하고 다시 SIARD 포맷에서 DBMS로 복원하는 과정을 통해 재현성을 확인하는 것이 목적이다. 국내 DBMS의 약 90%를 차지하는 3종 DBMS(MySQL, SQL Server, Oracle)가 검증 대상이며, 〈표 14〉처럼 4단계로 진행한다(행정안전부, 2019).

3종의 DBMS는 DB의 구조와 사용하는 Data Type이 다르므로 DBMS 제조사에서 제공하는 매뉴얼을 참고하여 최대한 많고 다양한 Data Type을 이용해 DB 생성하였다. 검증 시험에 사용한 3종 DBMS의 Data Type은 〈표 15〉와 같다.

SIARD 변환 및 복원 검증 시험 환경 및 SIARD 변환 및 복원을 수행하는 소프트웨어 정보는 〈표 16〉과 같다.

〈표 13〉 RDB형 데이터세트 보존포맷으로서 최종 적합성 평가결과: SIARD

구분	평가내용		합계
	공통기준	고유기준	
점수(총점)	22(33)	12(15)	34(48)
평점(100%)	67(100)	80(100)	74(100)
등급	D(보통)	C(양호)	C(양호)
최종 평가 결과	부분적합		

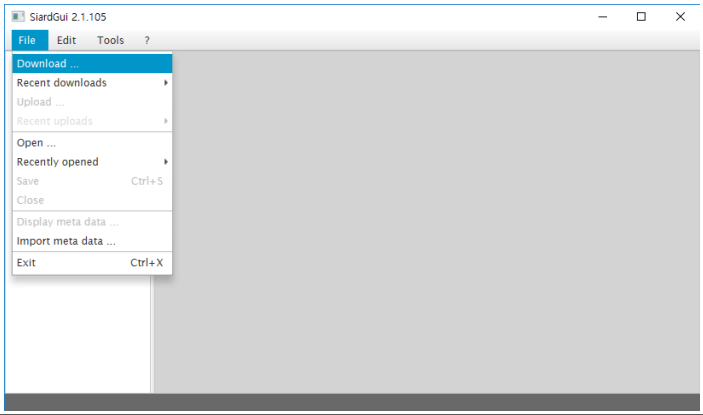
〈표 14〉 보존포맷 변환 및 복원 검증 시험 방법

순서	상세 내용
1. 원본DB 생성	• 3종의 DBMS에서 각각 DB 생성 • DB 생성 시 Routine Type도 포함하여 생성
2. (변환) DBMS에서 Download (DBMS → SIARD 파일)	• 생성한 DB를 SIARD 파일로 변환
3. (복원) 동일한 DBMS로 Upload (SIARD 파일 → DBMS)	• SIARD 파일을 본래의 DBMS로 Upload
4. 원본DB와 복원DB 데이터 확인	• Data, Key, Routine Type 보존 여부 확인

〈표 15〉 3종 DBMS Data Type

Data Type		DBMS 종류		
		MySQL 8.0	SQL Server 2017	Oracle 11g
일반 Data Type	숫자	BIT, INT, TINYINT, SMALLINT, MEDIUMINT, BIGINT, NUMERIC, DECIMAL, DOUBLE, REAL, FLOAT, BOOLEAN	BIT, INT, TINYINT, SMALLINT, BIGINT, MONEY, SMALLMONEY, NUMERIC, DECIMAL, FLOAT, REAL	NUMBER, FLOAT, BINARY_FLOAT, BINARY_DOUBLE
	문자/이진	CHAR, VARCHAR, BINARY, VARBINARY	CHAR, NCHAR, VARCHAR, NVARCHAR, BINARY, VARBINARY,	CHAR, VARCHAR2, NCHAR, NVARCHAR2
	대형 객체	BLOB, TINYBLOB, MEDIUMBLOB, LONGBLOB, TEXT, TINYTEXT, MEDIUMTEXT, LONGTEXT	TEXT, NTEXT, IMAGE	LONG, RAW, LONG RAW, BLOB, BFILE, CLOB, NCLOB
	날짜/시간	DATE, TIME, DATETIME, TIMESTAMP, YEAR	DATE, TIME, DATETIME, DATETIME2, DATETIMEOFFSET, SMALLDATETIME	DATE, TIMESTAMP, TIMESTAMP WITH TIME ZONE, TIMESTAMP WITH LOCAL TIME ZONE, INTERVAL YEAR, TO MONTH, INTERVAL DAY, TO SECOND
특수 Data Type		JSON, GEOMETRY, POINT, MULTIPOINT, LINESTRING, MULTILINESTRING, POLYGON, MULTIPOLYGON, GEOMETRY, COLLECTION, ENUM, SET	geography, geometry	ROWID, UROWID

〈표 16〉 보존포맷 변환 검증 시험 환경 및 SIARD 소프트웨어 정보

구분	내용
HW 스펙	CPU: i7-8750H 2.2GHz, RAM: 32GB, SSD: 1TB
OS 버전	Windows 10
SIARD 소프트웨어	Siard Suite 2.1.105 (SIARD Suite 다운로드: https://github.com/sfa-siard/SiardGui/releases) 

SIARD 변환 및 복원 검증 시험 결과는 〈표 17〉처럼 요약될 수 있다. 숫자, 문자/이진, 대형객체, 날짜/시간을 '일반 Data Type'으로, 나머지를 '특수 Data Type'으로 분류하였다. Table 간의 관계를 보여주는 PK, FK은 'Key Type'으로 분류하였으며 테이블 간에 임의로 관계를 설정하였고, 'Routine Type'도 임의로 생성하여 검증 시험을 진행하였다.

1) MySQL↔SIARD 변환 및 복원 시험 결과

일반 Data Type은 모두 변환 및 복원이 가능하며, 특수 Data Type는 'JSON'을 제외하고는 변환 및 복원이 가능하다. MySQL의 SIARD 매핑 결과는 〈표 18〉과 같다.

〈표 17〉 '3종 DBMS' ↔ 'SIARD' 변환 및 복원 검증 시험 결과 요약표

항목 \ DBMS	MySQL	SQL Server	Oracle
일반 Data Type (숫자, 문자/이진, 대형객체, 날짜/시간)	◎	◎	◎
특수 Data Type (기타)	○	◎	○
Key Type (PK, FK)	◎	◎	○
Routine Type (Stored Procedure)	X	X	X

(◎: 모두 변환 가능, ○: 부분 변환 가능, X: 변환 불가능)

〈표 18〉 MySQL ↔ SIARD Data Type 변환 및 복원 결과

종류		Data Type	
		MySQL	SIARD(SQL:2008)
일반	숫자	BIT	BOOLEAN
		INT	INTEGER
		TINYINT	SMALLINT
		SMALLINT	
		MEDIUMINT	INTEGER
		BIGINT	BIGINT
		NUMERIC	DECIMAL
		DECIMAL	
		DOUBLE	DOUBLE PRECISION
		REAL	
		FLOAT	FLOAT
		BOOLEAN	SMALLINT
	문자/이진	CHAR	CHARACTER
		VARCHAR	VARCHAR
		BINARY	BINARY
		VARBINARY	VARBINARY
	대형객체	TINYBLOB	
		BLOB	BLOB
		MEDIUMBLOB	
		LONGBOLB	
		TINYTEXT	VARCHAR
		TEXT	CLOB
		MEDIUMTEXT	
		LONGTEXT	
	날짜/시간	DATE	DATE
		TIME	TIME
		DATETIME	TIMESTAMP
		TIMESTAMP	
		YEAR	SMALLINT
특수		JSON	변환 불가
		GEOMETRY	CLOB
		POINT	
		MULTIPOINT	
		LINESTRING	
		MULTILINESTRING	
		POLYGON	
		MULTIPOLYGON	
		GEOMETRYCOLLECTION	
		ENUM	VARCHAR
		SET	

MySQL의 Key Type은 정상적으로 변환 및 복원되지만, MySQL의 Routine Type(Stored Procedures)을 SIARD로 변환을 할 경우, "routines" 카테고리에 Routine Type의 이름 정보만 변환되고, Routine Type의 이름 정보를 포함한 모든 정보가 누락 되는 것을 확인하였다(〈그림 5〉 및 〈그림 6〉 참조).

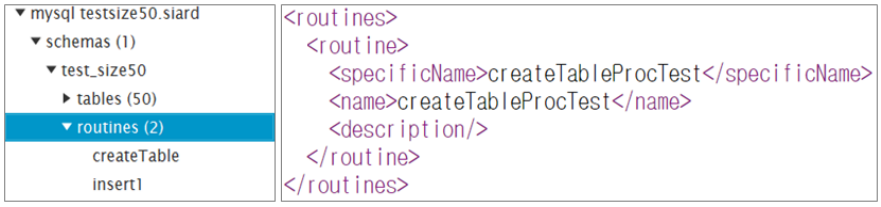
2) SQL Server ↔ SIARD 변환 및 복원 시험 결과

일반 Data Type, 특수 Data Type, Key Type 모두 변환 및 복원이 가능하다. MySQL의 SIARD 매핑 결과는 〈표 19〉와 같다. 단, Routine Type의 경우는 MySQL의 경우와 같은 이유로 정보가 누락되는 것을 확인하였다.

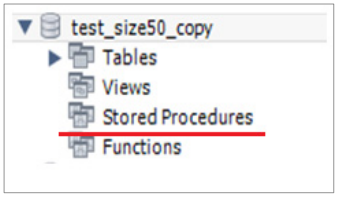
3) Oracle ↔ SIARD 변환 및 복원 결과

모든 일반 Data Type는 변환 및 복원이 가능하다. 단, 특수 Data Type의 중에서 "UROWID" 타입이 SIARD로 변환하는 도중에 에러가 발생하여 전체 변환 과정이 중단되고 SIARD Suite 소프트웨어가 강제 종료가 되었다. 해당 타입은 SIARD Suite에서 인식할 수 없는 타입으로 판단되며, DB 구조에서 'UROWID' 컬럼을 제외하면, 특수 Data Type도 모두 변환 및 복원이 가능하다(〈표 20〉 참조).

Key Type의 경우에는, PK와 FK 모두 SIARD로는 정상적으로 변환된 것을 확인하였다. 그러나 SIARD파일을 Oracle로 복원할 경우, PK는 정상적으로 변환이 되지만 FK는 누락되는 것을 확인하였다. FK의 경우, Oracle로 복원할 때 복원DB의 PK 제약조건 이름(Constraint Name)이 SIARD 파일에 있는 이름(〈그림 7〉의 제약조건 이름(SYS_C0012339) 참고)으로 설정되지 않고, DBMS(Oracle SQL Developer)가 임의로 변경(〈그림 8〉의 제약조건 이름(SYS_C0012347)하여 원본DB의 PK와 업로드DB의 PK를 동일한 것으로 인식하지 못하였다. 따라서 변경된 제



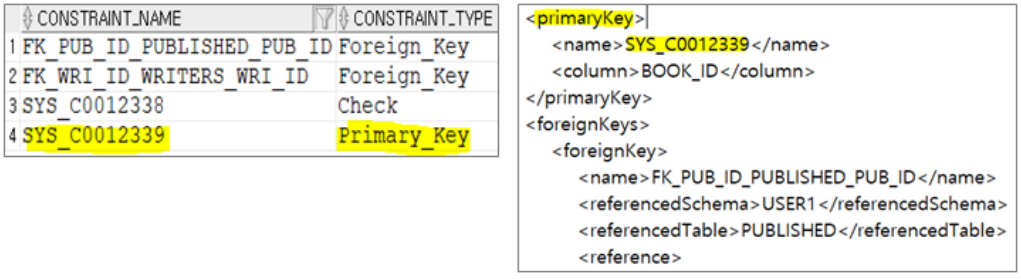
〈그림 5〉 MySQL → SIARD 파일 변환 후, SIARD Suite에서 Routine Type 확인



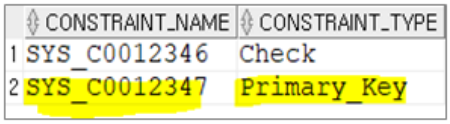
〈그림 6〉 MySQL로 복원 Workbench 화면

〈표 19〉 SQL Server ↔ SIARD Data Type 변환 및 복원 결과

종류		Data Type	
		SQL Server 2014	SIARD(SQL:2008)
일반 Data Type	숫자	BIT	BOOLEAN
		INT	INTEGER
		TINYINT	SMALLINT
		SMALLINT	
		BIGINT	BIGINT
		MONEY	DECIMAL
		SMALLMONEY	
		NUMERIC	NUMERIC
		DECIMAL	DECIMAL
		FLOAT	DOUBLE PRECISION
		REAL	REAL
	문자/이진	CHAR	CHARACTER
		NCHAR	NCHAR
		VARCHAR	VARCHAR
		NVARCHAR	NCHAR VARYING
		BINARY	BINARY
		VARBINARY(MAX)	VARBINARY
	대형 객체	TEXT	CLOB
		NTEXT	NCLOB
		IMAGE	BLOB
	날짜/시간	DATE	DATE
		TIME	TIME
		DATETIME	TIMESTAMP
		DATETIME2	
		DATETIMEOFFSET	VARCHAR
		SMALLDATETIMEOFFSET	TIMESTAMP
특수 Data Type		GEOGRAPHY	VARCHAR
		GEOMETRY	VARCHAR



〈그림 7〉 원본DB(왼쪽)와 SIARD파일(오른쪽)에 있는 제약조건 이름



	CONSTRAINT_NAME	CONSTRAINT_TYPE
1	SYS_C0012346	Check
2	SYS_C0012347	Primary_Key

〈그림 8〉 복원DB의 BOOK table 제약조건(PK, FK)

〈표 20〉 Oracle ↔ SIARD Data Type 변환 및 복원 결과

종류		Data Type	
		Oracle 11g	SIARD 2.1(SQL:2008)
일반 Data Type	숫자	NUMBER	DECIMAL
		FLOAT	FLOAT
		BINARY_FLOAT	REAL
		BINARY_DOUBLE	DOUBLE PRECISION
	문자/이진	CHAR	CHAR
		VARCHAR2	VARCHAR
		NCHAR	NCHAR
		NVARCHAR2	NCHAR VARYING
	대형 객체	LONG	CLOB
		RAW	VARBINARY
		LONG RAW	BLOB
		BLOB	
		BFILE	
		CLOB	CLOB
		NCLOB	NCLOB
	날짜/시간	DATE	DATE
		TIMESTAMP	TIMESTAMP
		TIMESTAMP WITH TIME ZONE	
		TIMESTAMP WITH LOCAL TIME ZONE	
		INTERVAL YEAR TO MONTH	INTERVAL YEAR TO MONTH
		INTERVAL DAY TO SECOND	INTERVAL DAY TO SECOND
특수 Data Type		ROWID	BIGINT
		UROWID	변환불가

약조건 이름을 가진 PK에 영향을 받아 FK는 누락되는 것으로 확인되었다. Routine Type의 경우는 MySQL, SQL Server의 경우와 같은 이유로 정보가 누락되는 것을 확인하였다.

4.3.3 재난안전정보 데이터세트의 SIARD 변환 및 복원 검증

임의로 생성한 DB의 데이터에 대한 SIARD 변환 및 복원 검증에 추가적으로 자체적으로 재난안전 관련 공공기관에서 공개한 재난안전 정보를 크롤링(Crawling)하여 자체 DB에 수

집한 실데이터를 대상으로 변환 및 복원 검증을 수행하였다. 〈표 21〉은 크롤링한 재난안전정보 DB에 대한 개요를 〈그림 9〉는 재난안전정보 DB의 ERD(Entity Relationship Diagram)을 각각 보여주고 있다. 실제로 재난관련 공공기관의 홈페이지에서 제공하고 있는 데이터들을 크롤링한 데이터는 대부분, INT, VARCHAR, DATETIME, TEXT, LONGTEXT, TIMESTAMP 등의 기본 Data Type으로 이루어져 있다.

재난안전정보 데이터세트의 SIARD 변환 및 복원 검증의 순서는 〈표 22〉와 같다.

검증 결과, 원본DB로부터 SIARD로 성공적으로 변환되었으며, SIARD에서 동일한 데이터베이스로 복원되어 복원DB가 생성되었다. 또한 TOAD Data Point를 이용하여 두 개의 스키마 사이의 데이터를 비교한 결과 모든 데이터가 동일하다는 것을 확인할 수 있었다(〈그림 10〉 참조).

5. 결 론

본 연구는 국내외 전자기록물 보존포맷 현황 및 선정 기준을 조사하였으며, 대규모의 행정정보 데이터세트를 생산·관리하고 있는 공공기관 행정정보시스템의 형태 및 운영 현황 등을 조사 및 분석하였다. SP(Significant Properties)를 통해 도출한 데이터세트의 주요 특성으로부

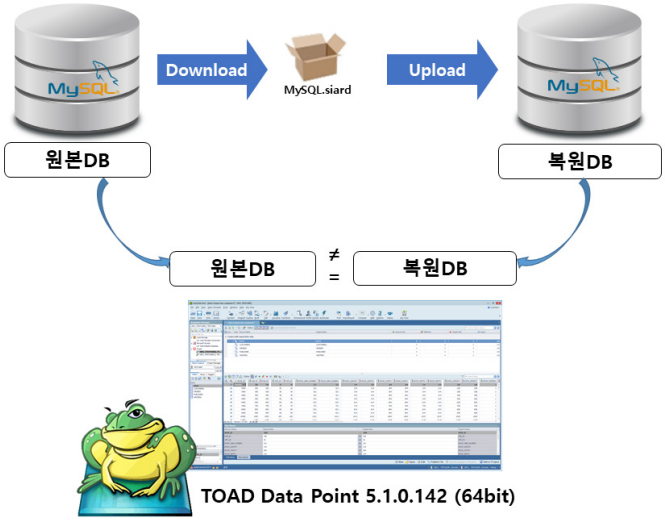
〈표 21〉 재난안전정보 DB 개요

항목	내용
이름	• 크롤링 DB
버전	• MySQL Ver 14.14 Distrib 5.7.26, for Linux (x86_64)
IP주소	• 113.198.***.***
Schema명	• crawling
Table개수	• 41개
등록건수	• 약 71만 건
크기	• 전체 약 704MB

〈표 22〉 보존포맷 변환 및 복원 검증 시험 방법

순서	상세 내용
1. 재난안전정보 크롤링 DB 생성	• 크롤링하여 재난안전정보 DB 생성 • Schema명crawling
2. (변환) DBMS에서 Download (DBMS → SIARD 파일)	• 원본DB를 SIARD 파일로 변환
3. (복원) 동일한 DBMS로 Upload (SIARD 파일 → DBMS)	• SIARD 파일을 DBMS로 Upload하여 복원DB 생성 • 업로드 전에 새로운 스키마 생성 • Schema명restore_crawling
4. 원본DB와 복원DB 데이터 비교	• TOAD Data Point를 이용하여 데이터 비교 • 데이터 비교 방법은 〈그림 10〉 참고

〈그림 9〉 재난안전정보 DB ERD(Entity Relationship Diagram)



〈그림 10〉 TOAD Data Point로 원본DB와 복원DB 데이터 비교

터 도출한 고유기준과 한희정, 오효정, 양동민(2020) 연구의 공통기준을 기반으로 데이터세트 유형 전자기록 보존포맷 선정을 위한 평가체계를 개발하였다. 그리고 국내외적으로 데이터세트 유형 전자기록의 보존포맷으로 채택 또는 검토되고 있는 SIARD 2.1을 대상으로 본 연구에서 개발한 평가체계를 적용하여 보존포맷으로서의 적합성을 검증하여 부분적합으로 판정하였고 국내 DBMS 소프트웨어 시장의 90% 이상을 차지하는 3종의 DBMS의 데이터 타입에 대해서 실험적으로 적합성을 평가하였다. 또한, 마지막으로 재난안전 관련 공공기관의 행정정보시스템에서 수집한 RDB형 데이터세트를 대상으로 SIARD로 변환하고 복원하는 검증 시험을 실시하였다.

SIARD는 RDB형 데이터세트의 내용뿐만 아니라 기능까지 보존하면서 표준 규격 및 오픈소스도 제공하므로 데이터세트 보존의 중요성이 커지고 있는 현재 시점에서 검토가 필요한 포맷이다. 본 연구에서의 평가결과 부분적합으로 판정되었기 때문에 보존포맷으로서 활용은 가능하지만 지속적인 검증이 필요하다고 판단된다. 데이터세트의 기능까지 보존되는 포맷으로는 상용 포맷인 Lindely(2013)의 CHRONOS 이외에는 SIARD가 유일하며, 유럽 E-ARK 프로젝트를 중심으로 지속적으로 개발되고 안정화되고 있으므로 지속적으로 모니터링이 필요하다. 국내에서는 DB에는 위치(서버 IP, 디스크 경로 등)만 저장하고, 해당 위치에 파일들을 저장하는 구조로 활용하고 있다. 이렇게 외부 위치(외부 디스크 등)에 저장되어 있는 파일들까지 보존해야 하는 필요성이 있으므로 덴마크의 SIARD-DK처럼 SIARD를 준수하면서 확장하는 방법도 고려할 필요가 있다.

본 연구를 통해서 데이터세트 유형 전자기록 보존포맷 선정을 위한 평가체계가 도출되었다. 이는 표준전자문서 중심의 국내의 단일 보존포맷 전략이 급변하는 기록관리 환경과 4차 산업

혁명 기술들에 유연하게 대응하고 시스템을 확장할 수 있는 정책적 기반을 마련한다는 점에서 의의를 갖는다. 그렇지만 향후 다음 3가지 측면을 지속적인 연구를 통해 보완할 필요가 있다. 첫 번째, 데이터세트 이외에 공공기록물법에 명시되어 있는 전자문서, 시청각기록물, 웹기록물, 간행물 등에 대해서도 각각의 전자기록 특성을 조사, 분석하여 고유기준 및 평가체계(항목, 정의, 설명, 가중치 등)를 도출해야 한다. 두 번째, 전자기록 보존포맷 선정위원회 또는 자문기구를 두어 정기적으로 보존포맷 선정 및 평가체계를 유지 및 관리해야 한다. 보존포맷은 IT기술의 발전에 영향을 받기 때문에 이를 보존포맷 선정 및 평가체계에 지속적으로 반영할 수 있는 시스템이 마련되어야 한다. 또한, 기준 항목에 대한 가중치, 보존포맷 선정평가에는 정량적인 평가 뿐만 아니라 정성적인 평가 결과도 포함되어야 한다. 그러므로 영구기록물관리기관에서는 정기적으로 운영되는 보존포맷 선정위원회 또는 자문기구 성격의 조직을 구성이 필요하다. 마지막으로, 데이터세트 유형 전자기록 보존포맷으로 선정된 파일포맷을 보존에 활용하기 위해서는 영구기록물관리기관이 주관하여 기록관 또는 기록원에서 참조할 수 있는 오픈소스코드(SFA의 SIARD Suite, RODA Database Preservation Toolkit 등)를 제공하는 것이 필요하다. 데이터세트는 다양한 DBMS 제조사가 존재하며 각 DBMS마다 독특한 기능과 데이터 타입을 제공하지만, 보존포맷은 모든 DBMS의 전체 기능과 데이터 타입을 지원하는 것은 불가능하다. 그러므로 영구기록물관리기관은 오픈소스코드 공개와 함께 새로운 DBMS 지원하여 확장하는 방안도 함께 제공할 필요가 있다.

참 고 문 헌

강현민 (2016). 중앙기록물관리기관의 종이기록물 영구보존용 마스터 파일로서 JPEG 포맷의 표준화에 대한 연구. 한국도서관·정보학회지, 47(4), 489-510. https://doi.org/10.16981/kliss.47.4.201612.489

국가기록원 (2004). 전자기록물 영구보존 기반기술 용역 완료보고서. 대전: 국가기록원.

국가기록원 (2013). 행정기관 전자기록물 재현기술 연구 및 프로토타입 개발 완료보고서. 대전: 국가기록원.

국가기록원 (2017). 차세대 기록관리 모델 재설계 연구 개발 완료보고서. 대전: 국가기록원.

노종원, 소정의 (2020). 데이터세트의 장기적인 보존 및 활용을 위한 관리 방안에 관한 연구. 디지털문화아카이브지, 3(1), 51-64.

박준영, 이명규 (2019). 디지털 사진기록물 관리를 위한 Raw 이미지 파일 포맷의 도입에 관한 연구. 한국기록관리학회지, 19(3), 155-178. https://doi.org/10.14404/JKSARM.2019.19.3.155

성환혁 (2007). 전자기록의 장기적 보존 및 활용을 위한 유형별 문서보존포맷에 관한 연구. 석사학위논문, 한국외국어대학교 대학원.

소정의 (2019). 데이터세트 보존포맷 선정을 위한 주요 항목 도출에 관한 연구 - 관계형 DB의 데이터세트를 중심으로. 석사학위논문, 전북대학교 대학원.

송치호, 차현철 (2017). 장기보존 전자기록의 위험평가에 관한 연구. 한국컴퓨터정보학회 동계학술대회, 25(1), 29-30.

오세라, 박승훈, 임진희 (2018). 행정정보 데이터세트 사례조사 연구. 한국기록관리학회지, 18(2), 109-133. https://doi.org/10.14404/JKSARM.2018.18.2.109

오세라, 이해영 (2019). 행정정보 데이터세트의 기록관리 방안. 한국기록관리학회지, 19(2), 51-76. https://doi.org/10.14404/JKSARM.2019.19.2.051

오세라, 정미리, 임진희 (2016). 공개포맷에 기반한 전자기록 보존 포맷 재설계 방향 연구. 한국기록관리학회지, 16(4), 79-120. https://doi.org/10.14404/JKSARM.2016.16.4.079

왕호성, 설문원 (2017). 행정정보 데이터세트 기록의 관리방안. 한국기록관리학회지, 17(3), 23-47. https://doi.org/10.14404/JKSARM.2017.17.3.023

임나영, 남영준 (2019). 기록의 디지털화 기준에 관한 연구. 한국비블리아학회지, 30(3), 5-30. https://doi.org/10.14699/kbiblia.2019.30.3.005

정부 (2017). 데이터기반행정 활성화에 관한 법률안. 의안번호 11077. 정부입법지원센터. Retrieved from http://www.lawmaking.go.kr

차현철, 최주호 (2019). 전자기록의 장기보존을 위한 위험평가 방법의 제안. 멀티미디어학회지, 22(1), 79-87. https://doi.org/10.9717/kmms.2019.22.1.079

한국기록학회 편 (2008). 기록학 용어 사전. 서울: 역사비평사.

한희정, 오효정, 양동민 (2020). 전자기록물의 장기보존을 위한 보존포맷 선정 방안에 관한 연구. 한국기록관리학회지, 20(1), 69-87. https://doi.org/10.14404/JKSARM.2020.20.1.069

행정안전부 (2018. 9. 19). 공공부문 원천데이터, 보존 의무화 된다. 보도자료. 행정안전부.

행정안전부 (2019). 2019년도 범정부EA기반 공공부문 정보자원 현황 통계 보고서. 대구: 한국정보화진흥원.

현문수 (2005). 데이터세트 기록의 관리 방안. 한국기록관리학회지, 5(2), 103-124. https://doi.org/10.14404/JKSARM.2005.5.2.103

eCH-0165 (2018). SIARD format specification. Version 2.1

eCH-0233 (2019). Archivierung elektronischer steuerdaten und -akten der kantone. Version 1.0

Essen M. V., Rooij, M. D., Roberts, B., & Dobbelsteen, M. V. D. (2011). Database preservation case study: Review. National Archives of the Netherlands.

Giaretta, D., Matthews, B., Bicarregui, J., Lambert, S., Guercio, M., Michetti, G., & Sawyer D.

(2009). Significant properties, authenticity, provenance, representation information and OAIS Information. Paper presented at the iPRES 2009: the Sixth International Conference on Preservation of Digital Objects, San Francisco, California. https://escholarship.org/uc/item/0wf3j9cw

Knight, G. (2008). Framework for the definition of significant properties. The National Archives, InSPECT Project Document.

Lindely, A. (2013). Database preservation evaluation report -SIARD vs. CHRONOS Preserving complex structures as databases through a record centric approach?. International Conference on Preservation of Digital Objects (iPres), Lisbon. https://doi.org/10.13140/2.1.3272.8005

NARA (2009). Significant properties. Retrieved from https://www.archives.gov/files/era/acera/pdf/significant-properties.pdf

The National Archives (2018. 5. 1). Significant properties. Retrieved from http://www.significantproperties.org.uk

• 국문 참고문헌에 대한 영문 표기
(English translation of references written in Korean)

Cha, H.-C., & Song, C.-H. (2019). A risk assessment method for the long-term preservation of electronic records. Journal of Korea Multimedia Society, 22(1), 79-87. https://doi.org/10.9717/kmms.2019.22.1.079

Han, H.-J., Oh, H.-J., & Yang, D. (2020). A study on the selection of preservation format for long-term preservation of electronic records. Journal of Korean Society of Archives and Records Management, 20(1), 69-87. https://doi.org/10.14404/JKSARM.2020.20.1.069

Hyun, M. (2005). A study on the management of dataset as records. Journal of the Korean Association of Records Management, 5(2), 103-124. https://doi.org/10.14404/JKSARM.2005.5.2.103

Kang, H. M. (2016). A study on the standardization of jpeg format as a long-term preservation master file for paper archives in the central archives of Korea. Journal of the Korean Library And Information Science Society, 47(4), 489-510. https://doi.org/10.16981/kliss.47.4.201612.489

Korea Minisry of Government (2017). Act on activation of data-based administration. Bill number 11077. Korea Ministry of Government Legislation. Retreived from http://www.lawmaking.go.kr

Korea Society of Archival Studies (2008). Dictionary of records and archival terminology. Seoul: Yuksa Bipyung Sa.

Lim, N., & Nam, Y. (2019). A study on the criteria for digitization of records. Journal of the Korean BIBLIA Society for library and Information Science, 30(3), 5-30. https://doi.org/10.14699/kbiblia.2019.30.3.005

Ministry of the Interior and Safety (2018). Source data of public sector, preservation is mandatory. Press release. 2018.09.19.

Ministry of the Interior and Safety, National Information Society Agency (2019). Statistical report on public sector information resources based on the EA in 2019.

National Archives of Korea (2004). Electronic record permanent preservation based technology service. Daejeon: National Archives of Korea.

National Archives of Korea (2013). A study on the reproduction technology and the prototype for the electronic records of administrative agency. Daejeon: National Archives of Korea.

Oh, S.-L., & Rieh, H.-Y. (2019). Managing data set in administrative information systems as records. Journal of Korean Society of Archives and Records Management, 19(2), 51-76. https://doi.org/10.14404/JKSARM.2019.19.2.051

Oh, S.-L., Jung, M. R., & Yim, J. H. (2016). Redesigning electronic records preservation formats based on open formats. Journal of Korean Society of Archives and Records Management, 16(4), 79-120. https://doi.org/10.14404/JKSARM.2016.16.4.079

Oh, S.-L., Park, S., & Yim, J. H. (2018). A case study of dataset records in information management system. Journal of Korean Society of Archives and Records Management, 18(2), 109-133. https://doi.org/10.14404/JKSARM.2018.18.2.109

Park, J., & Lee, M. (2019). A study on the introduction of raw image file formats for the management of digital photographic records. Journal of Korean Society of Archives and Records Management, 19(3), 155-178. https://doi.org/10.14404/JKSARM.2019.19.3.155

Roh, J.-W., & So, J.-E. (2020). A study on the management plan for preservation and long-term use of datasets. Journal of D-Culture Archives, 3(1), 51-64.

Seong, H. H. (2007). A study on document preservation format classified by the type for long-term preservation and use of electronic records. Master's thesis, Hankuk University of Foreign Studies. Seoul.

So, J. E. (2019). A study on derivation critical factor for selection of dataset preservation format: Focus on dataset of relational database. Master's thesis, Jeonbuk National University of Graduate School. Jeonju.

Song, C.-H., & Cha, H.-C. (2017). A study on the risk evaluation of electronic records for long-term preservation. Journal of The Korea Society of Computer and Information Winter Conference, 25(1), 29-30.

Wang, H.-S., & Seol, M.-W. (2017). A study on managing dataset records in government information systems. Journal of Korean Society of Archives and Records Management, 17(3), 23-47. https://doi.org/10.14404/JKSARM.2017.17.3.023