데이터세트 유형 전자기록의 필수보존속성 연구*

A Study on Significant Properties for Dataset Type Preservation Format

이 정 은 (Jung-eun Lee)**
양 동 민 (Dongmin Yang)***

초 록

본 연구는 이제까지의 전자기록 장기보존 정책이 문서유형 위주의 전자기록에 치중한 점과 다양한 행정정보시스템을 통해 생산되는 문서유형 이외의 전자기록 장기보존에 관한 문제 인식에서 시작되었다. 빅데이터 시대의 도래로 데이터 관리에 관한 관심이 높아지는 시점에서, 데이터세트를 장기적으로 보존하기 위한 고유기준 마련이 필요하다. 전자기록의 보존포맷은 해당 유형 전자기록의 고유기준에 의해 선정되며, 이 고유기준은 전자기록 유형에 따른 필수보존속성을 기준으로 마련된다. 이에 본 연구는 데이터세트 유형의 보존포맷선정 고유기준 마련에 앞서 데이터세트 유형의 전자기록에 관한 필수보존속성을 도출하는데 목적이 있다. 이를 위해 미국 NARA와 국가기록원이 수행한 R&D 연구결과를 비교・분석하였다. 연구의 결과로 데이터베이스형 필수보존속성 9개와 구조화데이터형 필수보존속성 7개를 도출하였다.

ABSTRACT

This study acknowledges that prevailing regulation concerning for the long-term preservation of electronic records focus mainly on document types, neglecting the preservation of electronic records from various administrative information systems. With the growing interest in data management in the era of big data, it is imperative to establish clear standards for the long-term preservation of datasets. The choice of preservation format for electronic records is based on the specific standards for each type of electronic record. These standards are formulated according to the significant properties relevant to the electronic record type. This study aims to identify the significant properties of electronic records of each record type, before creating specific preservation format selection criteria for these record types. To achieve this, we reviewed and analyzed R&D studies by the National Archives of Korea and the NARA in the United States. As a result of the research, 9 significant properties were identified for database-type entities, and 7 significant properties were identified for structured data-type entities.

키워드: 데이터세트, 필수보존속성, 장기보존, 보존포맷
Dataset, Significant Properties, Long-Term Preservation, Preservation Format

* 본 연구는 “2023년 행정안전부 국가기록원 기록관리 연구개발사업”의 연구비를 지원받아 수행되었음.
본 연구는 전북대학교 4단계 BK21 대학원혁신지원사업의 지원을 받아 수행된 연구임.
** 전북대학교 기록관리학과, 4단계 BK21사업단, 박사후 연구원(je.lee@jbnu.ac.kr) (제1저자)
*** 전북대학교 기록관리학과 부교수, 문화융복합아카이빙연구소 공동연구원(dmyang@jbnu.ac.kr) (교신저자)
논문접수일자 : 2023년 11월 21일 논문심사일자 : 2023년 11월 23일 게재확정일자 : 2023년 12월 8일
한국비블리아학회지, 34(4): 259-283, 2023. http://dx.doi.org/10.14699/kbiblia.2023.34.4.259

1. 서 론

1.1 연구 배경 및 목적

다양한 정보시스템을 통해 생산되는 데이터는 빅데이터 및 인공지능과 같은 기술의 진화를 촉진했다. 특히, 빅데이터라 명명되는 데이터들은 양, 속도, 다양성 측면에서 전례 없는 증가를 가져왔으며, 그 활용 분야 역시 기업 경영, 의료 서비스, 과학 연구 등 다양한 영역에서 사용되고 있다. 이러한 추이는 행정의 영역에서도 마찬가지이다. 예를 들어, 공공 보건과 관련한 위기 상황 발생 시 신속하고 정확한 데이터 분석은 적절한 대응 전략을 수립하는 데 필수적이며, 관련한 데이터가 잘 보존되지 않았을 경우, 위기 상황에 대한 반응이 지연되거나 부적합하게 처리될 수도 있다. 이처럼 공공기관에서 생산되는 데이터는 정책의 효율성과 정확성을 높이는 필수요소일 뿐만 아니라 의사결정을 가능케 하는 기반으로 작용하면서 데이터의 장기적 활용을 위한 보존의 중요성이 커지고 있다.

국가기록원은 이러한 시대적 흐름을 반영하여 2010년에 「공공기록물관리에 관한 법률(이하 공공기록물법)」 개정을 통해 공공기관의 행정정보 시스템에서 생산되는 데이터들을 '행정정보 데이터세트'라고 정의하였다. 이를 통해 데이터세트를 공공 영역에서 관리・보존해야 하는 전자기록의 한 유형으로 포섭하였으며, 이후에도 공공기록물법 시행령의 점진적인 부분 개정 및 기록관리 공공표준 등을 제정하면서 데이터세트에 대한 기록관리업무를 수행하고 있다.

한편, 전자기록은 파일이라는 매체를 통해 보존되는데, 기술 환경의 발전에 따라 파일 포맷의 종류나 특성이 다양해지고 이에 따라 생산되는 전자기록의 유형 역시 다양해지고 있다. 전자기록의 다양한 생산환경 변화는 2022년 공공기록물법 시행령의 개정에도 영향을 끼쳤다. 개정 전 시행령에서는 기록관의 전자기록물 보존과 관련하여 '문서보존포맷'이라는 용어를 사용하였으나, '문서'라는 단일 유형 전자기록 보존이라는 한계에 부딪히면서 해당 용어를 '보존포맷'으로 변경한 것이다. 국가기록원에서는 이러한 법령의 개정과 더불어 전자기록의 장기보존을 위한 보존포맷을 선정하고자 할 때 기술의 변화에 예속되지 않으면서도 다양한 유형의 전자기록 파일에 적용할 수 있는 기준을 마련하고자 하였다. 이에 2019년과 2020년에 연구개발 사업(이하 R&D 사업)을 수행하고, 그 결과물로 기록관리 공공표준인 「전자기록물 보존포맷 선정기준(v1.0), NAK 37:2022(v1.0)」을 제정하였다. 이 표준의 제정으로 2008년에 제정되었던 「전자기록물 문서보존포맷 기술규격, NAK 30:2008(v1.0)」 역시 「문서유형 전자기록물 보존포맷 기술규격: PDF/A-1b 기반의 포맷(v1.1), NAK 30:2022(v1.1)」로 변경되면서 문서유형에 해당하는 전자기록물의 보존포맷 기술 규격으로 내용이 수정되었다.

보존포맷은 전자기록물 생산 당시의 내용과 외형 등 주요 특성을 유지함으로써 시간과 기술의 변화에 상관없이 해당 기록물을 재현할 수 있는 포맷을 말한다(국가기록원, 2022a). 전자기록의 보존포맷과 관련한 기록관리 공공표준은 위에서 언급한 NAK 37:2022(v1.0)와 NAK 30:2022(v1.1)로 현재 2개가 제정되어 있다. NAK

37:2022(v1.0)에서는 모든 유형의 전자기록물에 적용되는 공통기준과 문서유형 전자기록물의 고유한 특성을 보존할 수 있는 고유기준을 제시하고 있다. 즉, 전자기록 유형의 관점에서는 전자문서 보존에 대한 정책만이 마련되어 있는 것이다.

서두에서 언급한 바와 같이 데이터세트 보존의 중요성은 날로 증대되고 있다. 문서유형의 보존포맷 선정 체계는 문서유형 전자기록물이 가지는 필수보존속성으로부터 도출되었다. 마찬가지로 행정정보 데이터세트의 장기보존포맷 선정을 위해서는 데이터세트가 가지는 필수보존속성의 도출이 선행되어야 할 것이다. 따라서 본 연구는 기록관리 관점에서의 데이터세트에 관한 필수보존속성을 도출하는 것을 목적으로 한다. 이를 위해 해외 국가기록원 중 유일하게 데이터세트에 관한 필수보존속성을 제시한 미국의 사례를 분석하고, 우리나라에 적용할 수 있는 필수보존속성을 제안하고자 한다.

1.2 선행연구

전자기록의 장기보존에 관한 연구는 전자기록관리라는 패러다임의 변화와 함께 2000년대 초반부터 꾸준히 연구되어왔다. 남성운, 윤대현(2001)의 연구는 초기 연구라고 할 수 있는데, 종이 기록물이라는 실체 보존에서 전자기록물이라는 보존 대상 변화에 따른 보존개념의 재정립과 장기보존을 위한 접근방법 등을 논의한 연구이다. 김명옥, 리상용(2010)은 전자기록의 장기보존을 위한 디지털 아카이브시스템 구축 시 필요한 기능요소를 제안하였다. 이를 위해 ISO 14721 OAIS 참조모형을 기반으로 기능요소를 도출한 후, InterPARES, Planets와 같은 국제 연구 프로젝트와 국내외 기록물관리기관의 디지털 아카이브시스템 구축 사례를 비교 분석하였다. 연구의 결과로 입수, 저장, 데이터관리, 보존계획, 운영, 접근의 6가지 영역에서 18개의 기능요소를 제안하였는데, 이 중 핵심요소를 저장, 데이터관리, 보존계획 영역으로 제시하였다. 특히 데이터관리 영역에서는 데이터베이스의 관리를 기능요소로 도출하였다. 소정의, 한희정, 양동민(2018)은 해외 국립아카이브 기관을 대상으로 전자기록물 장기보존 정책을 비교 분석하였다. 전자기록물의 장기보존 정책 요소를 보존 범위, 장기보존전략, 위험관리, 무결성 검증방식, 보존 인프라, 참조모델로 도출하였으며, 이 연구를 통해 해외 아카이브의 장기보존 범위 및 선호포맷, 허용포맷, 위험포맷 등의 개념이 소개되었다. 한희정, 오효정, 양동민(2020)은 국가기록원이 제시하는 장기보존포맷이 PDF/A-1으로 유일하다는 문제점을 개선하고자 연구를 진행하였다. IT의 발전으로 다양한 전자파일 포맷들이 업무에서 활용되고 있으나, 전자문서 이외의 유형에 대한 보존포맷 정책이 부재한 어려움이 있다. 이에 전자기록물의 장기적인 보존을 위한 보존포맷선정 방안을 고찰했고, 모든 전자기록물에 적용할 수 있는 보존포맷 선정기준인 공통기준과 평가 방식 등을 제안했다. 전한역 외(2023)는 시청각 유형 전자기록물의 보존포맷 선정을 위해 디지털 오디오를 중심으로 필수보존속성을 연구하였다. 영국, 미국의 국립아카이브에서 디지털 오디오가 어떻게 보존되고 있는지 분석하였으며, 오디오 유형의 특성에 따른 필수보존속성을 도출하였다. 특히, 이 연구에서는 영국의 국립기록보존소인 TNA의 필수보존

속성과 미국 국립문서기록관리청인 NARA에서 제안하는 필수보존속성의 범주를 분석하였는데, 분석의 결과로 우리나라 기록관리 공공표준에서 제시하는 필수보존속성 범주가 문서유형으로 한정된 점에 주목하였다. 이에 필수보존속성 관련 범주를 재정의하였으며, 본 연구에서도 새롭게 정의된 필수보존속성 범주를 참고하여 데이터세트 유형의 필수보존속성을 도출하고자 한다.

한편, 국가기록원이 주도하는 연구용역과제 중 전자기록의 세부 유형에 따른 연구는 2019년과 2020년에 진행되었다. 2019년의 연구는 데이터세트 유형 전자기록의 장기보존 기술과 관련한 연구이다. 연구의 결과로 데이터세트 유형 전자기록의 현황 및 해외의 장기보존기술이 조사되었으며, 데이터세트 보존포맷 선정기준 및 포맷관련 평가체계를 제안하였다. 해당 연구에서는 데이터세트 유형을 RDB(Relational DB)와 Non-RDB(Non-Relational DB)로 구분하였으며, 보존포맷 평가의 기준을 개방성, 안정성, 상호운용성, 자체 문서화, 검색기능으로 제시하였다(국가기록원, 2019). 이 연구를 통해 전자기록의 유형별 필수보존속성 개념이 소개되었으며, 본 연구의 대상이 되는 RDB형 데이터세트의 고유기준이 제시된 바 있다. 그러나 이 연구의 최종목적은 온전히 데이터세트 유형에 관한 보존포맷선정 관련 연구라기보다는 데이터세트 유형 전자기록의 장기보존을 위해 마이그레이션과 에뮬레이션 테스트베드를 구축하고, 데이터세트의 유형, 규모, 환경에 따른 실험을 통해 최적의 보존방식을 검증하는 것이었다. 따라서 본 연구에서는 이 연구보고서의 내용을 기반으로 후속연구를 진행하고자 한다. 2020년의 연구는 문서유형 보존포맷 및 장기보존패키지 다양화 연구이다. 텍스트형, 스프레드시트형을 비롯한 문서유형의 전자기록물 관리에 관한 연구로 문서유형의 전자기록 현황 및 보존방식에 대한 해외사례 분석이 진행되었으며, 문서유형 전자기록의 보존포맷 선정기준을 수립하고 권고포맷을 제시하였다(국가기록원, 2020).

선행연구를 검토한 결과, 전자기록의 장기보존과 관련한 연구는 보존정책에서부터 시스템, 장기보존포맷에 이르기까지 다양하였으나, 최근 들어서는 장기보존전략에 천착 됨을 알 수 있었다. 그런데도 여전히 문서유형 이외의 전자기록과 관련한 장기보존의 연구는 미미하다고 할 수 있다. 이에 본 연구는 데이터세트 유형의 전자기록을 연구의 대상으로 정하고, 장기보존전략의 기초 자료가 되는 데이터세트의 필수보존속성을 도출하고자 한다.

1.3 연구대상 및 방법

본 연구는 공공기관에서 생산되는 전자기록물 유형 중의 하나인 행정정보 데이터세트를 연구의 대상으로 한다. 우리나라의 법령에서 명시하는 정식적인 명칭은 '행정정보 데이터세트'이나 이후 서술에서는 데이터세트로 간략화하고자 한다. 본 연구의 목적은 데이터세트의 필수보존속성을 도출하는 것이며 연구 방법은 다음과 같다.

먼저 연구의 대상이 되는 데이터세트(Dataset)의 특징을 분석한다. 정보통신용어사전에 따르면, 데이터세트란 연관된 데이터를 모아서 특정 규칙에 따라 하나의 묶음으로 만든 데이터 집합을 말한다(한국정보통신기술협회, [발행년불명]).

이 데이터들의 집합은 데이터베이스라는 저장소에 저장되고, 기록의 재현을 위해서는 데이터베이스에서 불러내어져야 한다. 따라서 데이터베이스에 대한 특징 분석은 데이터세트의 필수보존속성을 도출하는데 선행돼야 할 것이다. 또한 기록의 관점에서 구조화데이터의 특징을 알아보고, 연구의 결과로 도출하게 될 필수보존속성의 개념을 알아본다.

데이터세트의 보존포맷 고유기준 및 파일포맷의 평가 방법을 도출하기 위해서는 데이터세트 장기보존 시 반드시 필요한 속성을 도출하는 것이 우선이다. 먼저, 문헌조사를 통해 미국, 영국, 호주 등의 해외 아카이브 현황을 조사하였는데, 대부분 국가들은 전자기록의 장기보존과 관련한 정책을 수립하고 있었으며, 전자기록물의 유형을 분류하고 유형에 따른 보존포맷 전략을 가지고 있었다. 전자기록물의 세부 유형 범주는 국가별로 약간의 차이를 보인다. 또한 이미지, 오디오, 비디오 유형의 전자기록에 대한 보존전략체계를 갖춘 국가들은 있었으나 데이터세트의 필수보존속성을 제시하고 보존포맷에 대한 정보를 공개하는 아카이브는 미국 국립문서기록청(이하 NARA)이 유일하였다. 이에 NARA에서 제시하고 있는 데이터세트의 필수보존속성을 분석하고자 한다.

마지막으로 2019년 국가기록원 R&D 연구보고서를 통해 제안된 데이터세트의 필수보존속성을 NARA의 필수보존속성과 비교・분석하여 우리나라에 적용할 수 있는 필수보존속성를 제안한다. 앞서 언급한 바와 같이 아직 우리나라에는 데이터세트의 장기보존을 위한 보존포맷 선정체계가 마련되지 않은 상황이다. 대신 2019년에 수행된 국가기록원 R&D 사업인 '데이터세트 유형 전자기록의 장기보존 기술 연구'에서는 전자기록 보존포맷 선정을 위한 공통기준과 함께 관계형 데이터베이스(Relation Database, RDB)형 데이터세트의 보존포맷 선정을 위한 고유기준이 도출된 바 있다. 따라서 이 연구의 후속 연구로써 현재 시점에서 보완되어야 할 내용 및 수정되어야 할 필수보존속성을 검토한다. 한편, 「기록관리 공공표준인 전자기록물 보존포맷 선정기준(NAK 37:2022(v1.0)」은 전자기록의 보존포맷을 선정하기 위한 공통기준과 문서유형 전자기록물에 대한 고유기준을 함께 제시하고 있다. 본 연구에서 도출된 필수보존속성은 데이터세트 유형 전자기록물의 보존포맷 선정 시 고유기준을 수립하는 기초자료로 활용될 수 있다. 따라서 도출된 데이터세트 유형 필수보존속성을 기반으로 고유기준으로 제시될 수 있는 필수보존속성의 정의 및 관련 설명을 제시한다.

향후 서술할 '데이터세트' 대한 용어는 같은 의미를 지니더라도 대응되는 언어의 표현에 따라 다르게 해석될 수 있다. 따라서 본 연구에서는 각 국가에서 사용된 언어를 그대로 사용하되 국내에 적용되는 대응어로는 'Structured Data: Database'는 '데이터베이스 유형'으로, 'Structured Data'는 '구조화데이터 유형'으로 통일하여 서술하겠다.

2. 이론적 배경

2.1 데이터베이스의 특징

데이터베이스는 전자적으로 저장되며 체계

적으로 관리되는 데이터세트를 말한다. 여기에는 단어, 숫자, 이미지, 비디오나 파일을 포함한 다양한 종류의 데이터가 포함될 수 있다. 데이터베이스는 보통 DBMS(Database Management System)라는 소프트웨어를 사용하여 데이터를 제어, 저장, 검색 및 편집하는 것이 특징이라고 할 수 있다. 같은 데이터세트 유형에 속하지만 '구조화데이터'와 '데이터베이스'의 가장 큰 차이점은 질의-응답 기능의 유무를 들 수 있는데, 구조화데이터에는 해당 기능이 없고, 데이터베이스에는 이 기능이 존재한다. 데이터베이스 중에서 현재 가장 많이 사용되고 있는 유형은 관계형 데이터베이스이며, 최근 비정형 데이터 처리 및 분석에 대한 요구가 높아지면서 NoSQL이 등장하였다. NoSQL은 해당 명칭에 대한 명확한 정의조차 내리지 못하고 있는 상황이며, 현재로서는 관계형 모델과 SQL(Structured Query Language)[1]을 사용하지 않으면서 데이터를 빠르게 처리하기 위해 내부적으로 트리, 그래프 등으로 데이터 구조를 활용하는 데이터베이스를 지칭한다(박우창, 남송휘, 이현룡, 2019).

데이터베이스에는 기본적으로 문자, 문자열, 정수, 실수, 시간, 날짜 등의 구조화된 데이터와 정형된 구조가 없는 텍스트 문서, 이미지, 동영상, 오디오 클립 등의 비정형 데이터가 저장된다. 비정형 데이터를 저장하기 위해서는 LOB(Large OBject)이라는 데이터 타입을 사용하는데 이들은 데이터베이스에 스키마라는 구조에 맞춰 저장·관리되는 특징을 가진다. 스키마에는 데이터 이외에 프로시저(Procedure), 함수(Function), 트리거(Trigger) 등과 같이 저장된 데이터에 대해 작업을 수행하는 명령문 형태의 루틴(Routine)을 내장할 수 있다. 관계형 데이터베이스에서 실제 데이터들이 저장되는 구조는 ERD(Entity Relationship Diagram)로 설계되며, ERD는 개체(Entity), 속성(Attribute), 관계(Relationship)로 표현된다(오세종, 2023).

데이터베이스는 그 규모가 증가할수록 기능이 다양해지므로 단순히 데이터만 저장되는 것이 아니라 간단한 계산, 특정 작업의 수행과 관련한 데이터세트의 요소가 증가하거나 복잡해지는 경향을 보인다. 따라서 원하는 계산이나, 작업 등을 수행할 수 있는 기능(프로시저, 함수, 트리거 등)을 프로그래밍을 통해 지원한다.

데이터세트에는 데이터세트와 연결된 파일이나 연계된 다른 데이터베이스 내의 데이터 등을 포함할 수 있다. 국내의 상당수의 정보시스템은 데이터세트 파일들을 데이터베이스에서 제공하고 있는 LOB에 직접 저장하지 않고, 외부에 저장한 후에 해당 파일의 경로를 데이터베이스에 저장하는 방식을 택하는 경우가 흔하다. 이러한 파일들은 해당 데이터세트의 핵심적인 부분을 차지하고 있는 경우가 많으므로 데이터세트를 정의하면서 반드시 포함되어야 한다. 연계 데이터베이스 내 데이터의 경우는 연계된 데이터를 참조만 하는지, 아니면 해당 데이터세트에 동기화 또는 복사되었는지에 따라 데이터세트의 범위에 여부를 결정하여야 한다. 단순히 값을 참조한 경우는 데이터세트를 정의함에 있어 포함할 필요가 없으며, 동기화 또는 복사된 경우는 해당 데이터세트에 포함되어야 할 것이다(박우창, 남송휘, 이현룡, 2019).

1) 관계형 데이터베이스의 데이터를 관리하기 위해 설계된 특수 목적의 프로그래밍 언어로, 데이터의 검색과 관리, 스키마 생성과 수정, 데이터 권한 관리 등을 위해 고안되었다.

DBMS의 측면에서의 특징을 살펴보면, DBMS에 따라서 제공하는 데이터, 스키마, 루틴 등이 다르므로, 구성요소의 구체적인 내용과 모습을 확인하기 위해서는 DBMS에 대한 정확한 정보 즉, 제조사, DBMS 명칭, DBMS 버전 등이 꼭 필요하다.

데이터베이스의 가장 중요한 기능은 질의-응답 기능이다. 사용자는 해당 데이터베이스로부터 응답받기 위해 미리 정해진 규칙과 문법을 사용해야 한다. 관계형 데이터베이스는 SQL(Structured Query Language)이라는 프로그래밍 언어를 공통으로 사용하여 질의-응답 기능을 제공하고, NoSQL은 각각의 데이터베이스마다 자체적으로 제공하는 API를 통해 질의-응답 기능을 지원한다. 사용자의 질문에 대한 답변은 다시 사용자 인터페이스 화면을 통해 제공되는데, 이 기능은 컴퓨터의 프로그래밍을 통해 구현된다. 즉, 질의-응답 기능의 확장된 형태로 볼 수 있다. 사용자 인터페이스는 데이터세트가 만들어진 본래의 목적을 확인할 수 있는 가장 중요한 기능이라고 할 수 있는데, 사용자 인터페이스의 형태는 대부분 웹브라우저에서 HTML 형식으로 제공되며, 내용은 업무 수행 과정・결과, 공문서, 증명서, 조회 화면 등 다양하게 제공된다(Elmasr & Navathe, 2015).

2.2 구조화데이터의 특징

구조화데이터(Structured Data)는 미리 정해 놓은 형식과 구조에 따라 저장되도록 구성된 데이터로 정의할 수 있다. 구조화데이터를 위해 미리 정해 놓은 형식과 구조는 사용자가 쉽게 이해하고 시스템에 쉽게 적용할 수 있도록 잘 알려진 포맷이나 명확한 데이터 구조 표현방법을 사용해야 한다. 구조화데이터의 대표적인 예는 관계형 데이터베이스의 테이블과 같이 고정된 컬럼에 저장되는 데이터와 지정된 행과 열로 데이터의 속성이 정해져 있는 스프레드시트(spreadsheet), 콤마로 구조가 결정되는 CSV 데이터 등이 있다(한국정보통신기술협회, [발행년불명]). 구조화데이터는 키-값으로 구성된 요소들을 2차원 테이블(CSV, TSV 등), 다차원 트리(XML, JSON 등) 등의 자료구조 방식으로 데이터들을 조직화할 수 있는 기능을 제공한다. 키-값으로 구성된 요소들은 예측할 수 있는 데이터 타입(정수, 실수, 문자, 문자열 등)을 가질 수 있으며, 이 구조화된 형태 덕분에 데이터를 검색, 분석, 수정 및 관리하는 것이 더욱 용이한 특징을 가진다.

XML은 HTML과 유사한 Mark-up 방식 즉, 태그(tag)로 데이터를 구분하여 구조화된 텍스트 파일 형태를 취하며, JSON은 키-값 쌍을 사용하여 데이터를 구조화한다. 구조화를 위한 구분기호는 큰따옴표(“ ”), 콜론(:), 쉼표(,), 중괄호({}), 대괄호([])를 활용하는 텍스트 파일을 사용한다. CSV는 구조화데이터 유형의 가장 대표적인 형식으로 각 데이터가 쉼표로 구분된 텍스트 파일로 구성된다. TSV는 각 데이터가 탭(Tab)으로 구분되어 있는 텍스트 파일 형식을 띤다. 앞서 살펴본 네 가지 형식의 파일은 모두 텍스트 파일로 텍스트 형식으로 표현할 수 있는 문자를 표시할 수 있다. 이렇게 문자로 표시할 수 있는 컴퓨터의 기본 단위는 0과 1로 표현되는 이진수이고, 이진수로 다양한 문자를 표현하는 방식을 문자 인코딩(Character Encoding)이라고 한다. 문자 인코딩 방식은 문

자를 비트의 조합으로 매핑하는 고유한 방법을 제공하며, ASCII, UNICODE의 문자 인코딩 표준 방식을 토대로 실제 인코딩된 비트를 저장하고 교환하기 위한 인코딩 체계로 UTF-8, ANSI가 대표적으로 쓰인다. 모든 텍스트 파일의 내용을 이해하기 위해서는 반드시 인코딩 방식에 대한 정보가 필수적이다. 또한, 문자 방식으로 인코딩할 수 없는 이미지, 음성, 영상 등의 이진 파일들도 Base64 인코딩 방식으로 텍스트 파일에 내재화시킬 수도 있다.

구조화데이터 형식에서 해당 데이터의 의미를 알기 위해 데이터를 구분하고 조직화하는 구조를 이해하고 잘 보존하여야 한다. 데이터 구조의 이해를 통해서 각 데이터가 어떻게 서로 관련되어 있는지를 명확히 이해할 수 있다(나시다 케이스케, 2018). 예를 들어, CSV 파일에서 쉼표는 서로 다른 데이터 필드를 분리하고, JSON에서는 중괄호와 대괄호가 객체와 배열의 시작과 끝을 나타냄을 이해할 수 있을 것이다. 이러한 구조에 대한 이해가 없거나 구조가 제대로 관리되지 못한다면 데이터의 의미를 해석하기에 어려울 수 있다. 최근 인공지능과 빅데이터의 활용이 증가하면서 구조화 데이터의 유형은 데이터 처리 및 분석에 가장 많이 사용되고 있는 파일포맷의 한 유형이라고 할 수 있다. 컴퓨터 프로그램과 알고리즘은 이러한 파일포맷을 사용하여 데이터를 파싱하고, 읽고, 쓰고, 처리하는데 예를 들어, XML 파서는 태그를 기반으로 데이터를 해석하고, CSV 파서는 쉼표를 기준으로 각 필드를 구별한다. 이때 만약 기호가 없거나 잘못된 위치에 있다면 데이터 파싱 과정에서 오류가 발생할 수 있다.

구조화데이터 유형 파일포맷은 데이터를 조직화하여 작성하는 고유한 문법이 존재하며, 해당 파일 객체가 문법을 준수하는지를 판단하는 것은 중요하다. XML 표준의 경우에는 작성 문법에 오류가 없는 문서를 'Well-Formed'라 명명하고, 'Well-Formed' 인지 아닌지를 판단하는 기능을 따로 두고 있다. 또한 자체 문서에 대한 Well-Formed 이외에도 XML, DTD, Schema 등과 같은 특정 양식을 정의하고 해당 양식에 대한 유효성을 검증하는 기능까지도 표준에 포함하고 있다. 데이터베이스와 마찬가지로 구조화데이터 역시 데이터세트에 연결된 데이터 객체가 존재할 수 있다. 따라서 Base64 인코딩 방식을 내재화시킬 수도 있지만, 데이터 객체를 외부에 저장한 후에 해당 파일의 경로를 구조화데이터 파일에 저장하는 방식을 택하고 있다.

2.3 전자기록의 필수보존속성

전자기록물의 필수보존속성(Significant Properties, SP)은 영국 국립기록보존소(이하, TNA)가 프로젝트 파트너로서 참여한 InSPECT (Investigating Significant Properties of Electronic Content) 프로젝트에서 처음 논의된 개념이다. InSPECT 프로젝트는 변화하는 기술 환경 속에서 시간의 경과에 따른 디지털 객체에 대한 접근성과 진본성 유지를 위한 연구로 이 프로젝트를 통해 전자기록의 유형 중 이미지와 오디오 포맷 중심의 필수보존속성이 도출되었다. InSPECT 프로젝트가 정의하는 필수보존속성은 '시간이 지나도 기록물에 접근할 수 있고, 기록물이 의미 있는 상태를 유지하기 위해 보존되어야 할 디지털 기록의 속성'으로 정의하고 있다(InSPECT, 2009). 전자기록을 장기

보존해야 하는 기관의 처지에서는 개체의 모든 요소가 아닌 보존을 위한 필수요소를 중심으로 효율적인 보존전략을 채택할 수 있다는 장점이 있다.

전자기록의 다양한 유형에 적합한 필수보존속성은 해외 각국의 아카이브에 의해 연구되고 있다. 현재까지 공통으로 도출된 필수보존속성은 외관(Appearance), 기능(Behavior), 내용(Content), 맥락(Context), 구조(Structure)를 포함한 5가지 범주로 분류된다(국가기록원, 2022a). 이 중에서 외관 범주와 관련된 용어는 미국 NARA는 'Appearance'를 사용하거나 영국 TNA는 'Rendering'을 사용하는 등 각국의 아카이브 방침과 관점에 따라 다른 용어를 사용하거나 병기하기도 한다. 두 용어 모두 전자기록의 재현에 관여하는 속성이라는 점에서 공통점을 가지나 미국 NARA에서는 '기록의 내용을 재현하고, 내용의 의미를 제공하는데 필요한 시각적 속성'으로 정의하면서 시각적 속성을 중시하는 경향을 보이고 있다. 반면, 영국 TNA에서는 '기록의 재현에 기여하는 모든 정보'로 정의되고 있어 시각적 표현 정보뿐만 아니라 청각적 표현정보 등에 필요한 모든 정보를 지칭하여 Appearance 정의보다 상위의 개념으로 이해된다. 우리나라 국가기록원에서도 이 다섯 가지 필수보존속성의 개념을 도입하여 문서유형 전자기록물의 보존포맷 선정 시 고유기준이 되는 기본 개념으로 활용하였다. 기록관리 공공표준 NAK 37:2022(v1.0)에서 정의하고 있는 필수보존속성의 주요 범주와 의미는 〈표 1〉과 같다.

3. 데이터세트의 필수보존속성 비교 분석

이 장에서는 먼저 해외 아카이브의 데이터세트 보존정책과 전자기록의 유형 범주를 분석하고, 이후 국가기록원이 시행한 2019년 R&D 보고서(관계형 데이터베이스)와 미국 NARA에서 제시하는 데이터세트 필수보존속성을 비교·분석하고자 한다. NARA에서는 데이터세트 유형의 디지털 기록을 총 4가지로 구분하고 있으나, 이번 연구에서는 데이터베이스 유형과 구조화데이터 유형으로 연구의 범위를 제한하고자 한다.

3.1 해외 아카이브의 보존정책 및 데이터세트 유형 정의

미국 NARA은 전자기록의 유형을 총 16개로 구분하고 있으며, 이 중 데이터세트 유형 전자

〈표 1〉 필수보존속성의 주요 범주(국가기록원, 2022a)

범주	의미
외관(Appearance)	기록의 외형적인 모습(예. 폰트, 색상)
기능(Behavior)	기록과 연결되어있는 외부와의 상호작용에 의한 기능
내용(Content)	기록 내 모든 데이터 및 수식
맥락(Context)	기록의 메타데이터(예. 작성자, 작성일)
구조(Structure)	기록의 구조정보

기록은 Structured Data:Calendars, Structured Data:Database, Structured Data, Structured Data:Spreadsheets로 총 4개가 제시되고 있다. 각 유형의 정의를 살펴보면 먼저 Structured Data:Database는 다양한 목적으로 조작, 검색, 추출할 수 있는 정보로 조직화, 구조화되어 있으며, DBMS를 통해 접근할 수 있는 정보로 정의하고 있다. Structured Data는 일반 텍스트로 구분되거나 마크업(Mark-up) 방식으로 구조화된 데이터를 말한다. 여기에는 테이블, 행, 열 등 명확한 구조를 가진 모든 종류의 데이터가 포함된다고 볼 수 있다. 따라서 관계형 데이터베이스, 스프레드시트 또는 마크업 텍스트 등을 포함할 수 있으며 데이터가 어떤 필드(fields) 또는 태그(tags)에 저장될 것인지를 설명하는 데이터 모델의 획득이 중요하다고 할 수 있다. 또한 숫자, 통화, 알파벳, 이름, 날짜, 주소 등과 같은 데이터의 유형과 통제어휘(controlled vocabulary)도 명시되어야 한다. 데이터베이스는 데이터베이스시스템(DBMS)이라는 특수한 소프트웨어를 통해 데이터를 관리하지만, 데이터베이스형이 아닌 구조화데이터세트는 일반 텍스트 파일에 담겨 있어 워드 프로세서 또는 메모장 등으로 확인할 수 있는 형식을 취한다. Structured Data:Calendars는 개인정보 관리든 조직 스케줄링 관리든 일반적으로 일정관리, 예약, 해야 할 목록 등 애플리케이션 내에 저장된 정보를 캡처하고 교환하는 데 필요한 구조화된 정보를 말한다. 캘린더는 기본적인 텍스트 형식 또는 디렉터리 구조의 MIME(Multipurpose Internet Mail Extensions) 유형이라고 할 수 있다(NARA, 2022c). 즉, 캘린더 데이터는 저장되고 전송되는 방식이 일반 텍스트 파일이나 디렉터리 구조를 기반으로 하되, 이를 캘린더 정보에 특화된 방식으로 확장하여 사용한다. Structured Data:Spreadsheets는 격자 모양의 행과 열로 배열된 데이터를 조작하고 공식에 따라 실행할 수 있는 전자문서를 말한다. 스프레드시트 소프트웨어는 일명 통합문서라 불리는 여러 시트를 허용할 수 있으며, 데이터를 텍스트, 숫자, 기호 또는 그래픽 형식으로 표시할 수 있다. 각 셀에는 원시 데이터 또는 외부 데이터 소스뿐만 아니라 다른 페이지 또는 시트의 다른 셀 내용을 기반으로 값을 자동으로 계산하고, 표시하는 수식의 결과를 포함할 수 있다(NARA, 2022d). NARA의 전자기록유형을 살펴본 결과, 본 연구와의 비교 대상이 되는 데이터는 Structured Data:Database, Structured Data임을 확인할 수 있다.

영국 국가기록원(이하 TNA)은 2011년부터 전자기록 보존에 대한 정책을 수립하였다. 2011년에 '디지털 보존정책: 아카이브즈를 위한 지침(Digital Preservation Policies: Guidance for archives)'을 발표하였고, 2017년에는 '디지털 전략(Digital Strategy)'을 중심으로 전자기록 장기보존과 관련한 내용을 보완하는 정책을 제시하였으며, 최근에는 '전략적 우선순위: 2023-27(Strategic priorities 2023-27)'을 통해 기록 보존소의 가치 및 영향을 광범위하게 극대화하기 위한 전략을 제시하였다(TNA, 2023). 특히, 2017년에 계획된 디지털 전략을 통해서는 이전의 전자기록 유형을 문서, 이미지, 이메일, 비디오, 웹사이트 등에 집중했으나, 구조화된 데이터세트와 컴퓨터 코드까지 포함하는 등 보존 범위를 확대하였다(TNA, 2017). TNA는 홈페이지를 통해 이관을 위한 파일 형식(File formats for transfer)

과 기관이 장기적으로 보존할 수 있는 디지털 파일포맷의 유형과 범위를 제시하고 있으며, 관련한 포맷의 명칭과 파일 확장자 등을 주기적으로 업데이트하고 있다(TNA, 2023). 이를 통해 전자기록의 유형을 살펴보면, 텍스트 기반 문서(Word processing and text), 스프레드시트(Spreadsheet), 프레젠테이션(Presentations), 이미지(Graphic), 오디오, 비디오, 이메일이 있으며, 유형별 보존 속성은 명시되어 있지 않다. 구조화데이터의 파일 확장자로는 ebcdic, xml, json, tsv, tab 등이 제시되었다.

호주 국립기록원(이하 NAA)에서는 디지털 기록을 디지털로 생산된 기록과 디지털화된 기록으로 구분하고 있다. NARA, TNA와는 달리 전자기록의 세부 유형 및 장기보존을 위한 필수요건들이 제시되고 있지는 않다. 다만, 원본 기록을 디지털화할 때, 원본 기록의 필수적인 특성(essential characteristics)을 보존하기 위한 규격사양(NAA, 2022)을 제시하고 있는데, 이를 통해 전자기록물의 유형 및 선호 포맷을 확인할 수 있다. 이를 기반으로 전자기록의 유형을 분류하면 업무 문서(Office documents), 이메일(Emails), 디지털 정지 이미지(Digital still image), 디지털 오디오(Digital audio), 디지털 시네마(Digital cinema), 디지털 비디오, 디지털 비디오 컨테이너(Digital Video (Container)), CAD(Computer Aided Design), 지리 공간(Geospatial), 구조화데이터(Structured Data), 웹사이트(Website)로 구분할 수 있다. 각 전자기록 유형에 대한 정의는 명시되어 있지 않으며, 데이터세트의 유형으로는 구조화데이터가 있다. NAA는 일관되고 체계적인 파일포맷 관리를 위해 선호(preferred) 포맷을 제시하고 있는데 구조화데이터의 포맷은 Comma Separated Value file, Extended Binary Coded Decimal Interchange Code, eXtensible Markup Language, JavaScript Object Notation, Tab Separated Value file이 제시되었다.

3.2 NARA의 데이터세트 필수보존속성

미국 NARA의 데이터세트 유형 전자기록은 모두 구조화데이터를 기준으로 하고 있으며, 우리나라의 데이터세트 유형과의 비교를 위해서는 Structured Data:Database, Structured Data가 비교 대상이 된다. NARA의 Structured Data:Spreadsheets 유형은 국가기록원의 분류 기준에 의하면 문서유형에 해당한다. 따라서 본 장에서는 Structured Data:Database 유형과 Structured Data 유형의 필수보존속성을 분석하고자 한다. TNA가 전자기록의 필수보존속성 범주를 5개로 정의한 반면, NARA는 내용(Content)을 제외한 외관(Appearance), 구조(Structure), 기능(Behavior), 맥락(Context)의 범주로 필수보존속성을 제시하고 있다. NARA(2009)에 의하면, 내용은 기록물에서 항상 중요한 것으로 여길 수 있는 속성 중 하나이기에, 아카이브 담당자가 이런 속성에 대해 추가적으로 판단할 필요가 없다. 이에 NARA는 내용을 제외한 4가지 유형의 필수보존속성을 제시하였다. 또한 의미상으로 내용은 기록물이 가진 내용물(Content) 자체를 뜻하기에, 오디오 같은 시청각 기록물의 고유한 속성을 분류하고, 이를 판단하기 위한 범주(도구)로서 효과적이지 않을 수 있다(전한역 외, 2023). 이에 본 연구에서는 NARA의 접근법과 유사한 맥락에서, 내

용(Content)을 필수보존속성 범주에서 제외하고자 한다.

TNA의 '내용' 범주의 의미는 '지적 작업물의 표현을 설명하는 추상적인 용어로 디지털 환경에서의 내용은 텍스트, 정지 이미지, 동영상, 오디오 등과 같은 지적 표현'이라고 명시하고 있다. 따라서 데이터세트는 고유한 지적 작업물이라기 보다는 데이터들의 재구성으로 언제든지 새로운 내용이 만들어질 수 있으므로 필수보존속성에서 제외된 것으로 판단된다.

Structured Data:Database의 범주별 필수보존속성을 살펴보면, 먼저 외관(Appearance)범주에서의 보존속성으로는 문자 인코딩(Character Encoding)과 텍스트 속성(Text Property)을 제시하고 있다. 단, 데이터베이스 유형에서 외관 범주는 반드시 필요한 범주에는 해당되지 않는다. 문자 인코딩이란 사용자가 입력한 글자나 기호들을 컴퓨터가 이용할 수 있는 신호로 만드는 것으로 사용자가 입력한 문자, 문자열, 정수, 실수, 날짜, 시간 등의 표현정보를 의미하며, 텍스트 속성은 글꼴, 텍스트 크기, 색상 등을 의미한다. 이러한 속성들은 현재는 주요한 속성으로 간주되지 않으나, 언제든 기록을 이해하는데 중요한 요소로 작용한다면 다시 고려되어야 할 속성이라고 명시하고 있다.

구조(Structure) 범주의 필수보존속성은 데이터베이스 스키마(Database Schema)와 기술 메타데이터(Technical Metadata)가 제시되었다. 데이터베이스 스키마란 데이터베이스에서 데이터의 구조와 그 표현법, 자료 간의 관계를 형식 언어로 정의한 것으로 데이터베이스 전체 또는 일부의 논리적인 구조를 표현한다. 즉, 데이터베이스 내에서 데이터가 어떤 구조로 저장되어 있는지를 나타내는 속성으로 데이터베이스 유형의 전자기록을 보존할 때 반드시 필요한 속성이라고 할 수 있다. 기술 메타데이터는 일반적으로 파일 헤더에 자동으로 내장되는 메타데이터로 특정 데이터베이스 형식이나 소프트웨어, 소프트웨어 버전 등을 설명해준다. 차후 데이터 활용을 위해 사용자와의 질의-응답 기능, 그래프 작성 등 데이터와의 상호작용을 위한 DBMS 정보를 의미한다.

기능(Behavior) 범주의 필수보존속성은 조작기능(Manipulation Functionality)과 질의 표시(Display Query)가 제시되어 있다. NARA는 데이터베이스 유형의 장기보존에 있어 기능 범주의 속성이 가장 중요하며, 그다음으로 구조 범주의 속성이라고 명시한다. 데이터베이스의 레코드는 어떤 종류의 데이터가 어떤 필드, 열 또는 태그에 저장될지를 설명하는 문서화된 데이터 모델이 필요하며, 구조에 대한 문서화는 결국 데이터베이스의 기능을 가능하게 하기 때문이다. 조작 기능은 테이블 내부 또는 테이블 간의 관계를 검사하는 기능을 말하며, 질의 표시는 질의 결과를 특정 형식으로 표시하는 것으로 질의 결과는 보고서, 조회 화면, 업무 수행 과정 및 결과 등이 있을 수 있다. 만약 질의 결과가 보고서라면 확장자가 .pdf, .doc와 같은 텍스트 파일이거나 지리정보라면 그래프와 표일 수도 있다.

맥락(Context) 범주의 필수보존속성은 시리즈(Series)와 설명 메타데이터(Descriptive Metadata)가 제시되었다. 시리즈는 서로 관련된 레코드의 그룹으로, 일반적으로 유사한 업무 활동이나 기능과 연관하여 하나의 단위로 사용되고 파일링되는 데이터를 말한다. 따라서

레코드 간의 관계와 그것들이 속한 시리즈가 보존되었을 때 기록의 맥락을 파악할 수 있을 것이다. 데이터베이스는 다수의 업무를 위해 만들어졌으며, 각 업무는 하나의 데이터세트에 대응되므로, 하나의 데이터베이스는 여러 개의 데이터세트로 구성될 수 있다. 설명 메타데이터는 레코드 내에 포함되는 정보로서, 자료의 지적 내용을 참조하고, 원하는 자료를 찾는 데 도움이 되는 정보를 말한다. 설명 메타데이터는 일반적으로 데이터베이스 자체의 구조와 내용에 포함되어 있다고 가정하며, 제목, 주제, 날짜, 이벤트, 생산자 등이 포함될 수 있으며, 이 외에도 여러 설명 메타데이터가 있을 수 있다.

Structured Data의 필수보존속성은 데이터베이스와 마찬가지로 외관, 구조, 기능, 맥락으로 제시되고 있다. 앞서 살펴본 바와 같이 데이터베이스 유형에서는 외관 범주가 필수보존속성으로 제시되지 않았다. 반면, 구조화데이터 유형에서는 문자 인코딩을 필수보존성으로 제시하고 있다. 이는 구조화 데이트의 정의에서 비롯된 것으로 유추된다. 구조화데이터는 레코드나 파일 내의 고정된 필드에 저장되는 모든 데이터로 정의되고 있는데, 여기에는 관계형 데이터베이스를 포함한 스프레드시트, 마크업된 텍스트에 포함된 데이터 등을 포함할 수 있기 때문이다. 문자 인코딩은 항상 존재해야 하는 속성이며, 호환 가능한 형식으로 열거나 ASCII와 같이 또 다른 포맷으로 변환하기 위해 반드시 식별되어야 함을 전제로 하고 있다.

구조(Structure) 범주에서는 데이터베이스에서 제시되었던 스키마와 기술 메타데이터를 포함하여 연결(Linkage), 열 개수Column Count), 행 개수(Row Count) 등이 추가로 제시되고 있다. 연결은 레코드, 파일 사이 혹은 내부적인 관계성을 의미하며, 열 개수와 행 개수는 문서 내의 전체 열과 행의 개수를 의미한다.

구조화데이터 기능 범주의 핵심은 하이퍼링크(Hyperlinks)이다. 하이퍼링크는 파일 내·외부 또는 외부 데이터 원본에 대한 링크를 의미하므로 NARA가 명시하는 기능 범주에 대한 정의, 즉 기록과 상호작용할 수 있는 특징을 반영한 필수보존속성이라 할 수 있다. 하이퍼링크의 가장 큰 위험은 해당 하이퍼링크가 시리즈 일부가 아닐 수 있는 외부 파일에 대한 링크이거나 더 이상 유효하지 않은 외부 웹사이트로의 링크인 경우이다.

맥락(Context) 범주의 필수보존속성은 스프레드시트에서 참조할 수 있는 관련된 파일이나 연결된 파일로써의 관련 파일(Related Files)이 있다. 이상으로 데이터베이스 유형 및 구조화데이터 유형의 필수보존속성 및 각 유형에 제시된 선호포맷과 허용포맷을 정리하면 〈표 2〉와 같다.

3.3 국가기록원과 NARA의 필수보존속성 비교분석

국가기록원은 2019년 R&D 연구사업을 통해 데이터세트 유형 전자기록 중 관계형 데이터베이스에 대한 필수보존속성을 도출한 바 있다. 해당 연구에서 제안한 필수보존속성 역시 InSPECT 프로젝트에서 제시하는 외관, 기능, 내용, 맥락, 구조에 해당하는 범주를 기준으로 필수보존속성을 도출하였다. 따라서 본 장에서는 앞서 분석한 미국 NARA의 필수보존속성과 국가기록원 R&D 연구결과를 중심으로 데

〈표 2〉 NARA의 Structured Data:Database 및 Structured Data의 필수보존속성

범주	필수보존속성	
	Structured Data:Database	Structured Data
외관 (Appearance)	• 문자 인코딩(Character Encoding) • 텍스트속성(Text Properties)	• 문자 인코딩(Character Encoding)
기능 (Behavior)	• 조작 기능(Manipulation Functionality) • 질의 표시(Display Query)	• 하이퍼링크(Hyperlinks)
구조 (Structure)	• 데이터베이스 스키마(Database Schema) • 기술 메타데이터(Technical Metadata)	• 스키마(Schema) • 연결(Linkage) • 열 개수(Column Count) • 행 개수(Row Count) • 기술 메타데이터(Technical Metadata)
맥락 (Context)	• 시리즈(Series) • 설명 메타데이터(Descriptive Metadata)	• 관련 파일(Related Files)
선호포맷	• Comma Separated Value(CSV) • OpenDocument Format Spreadsheet(ODS) • ASCII Text	• Comma Separated Value(CSV) • ASCII Text • XML • JSON • OpenDocument Format Spreadsheet
허용포맷	• Microsoft Excel Office Open XML • Microsoft Excel 97 Binary Document Format	• EBCDIC • Microsoft Excel Office Open XML • Microsoft Excel 97 Binary Document Format

이터세트 유형의 필수보존속성을 비교하였다. 분석의 결과로 국가기록원 필수보존속성의 특징은 다음과 같이 정리할 수 있다.

첫째, 필수보존속성의 범주인 외관(Appearance)을 제외하였다. 이는 데이터세트를 외관이라는 정의에 비추어 판단하였을 때, 전자문서와는 달리 글꼴, 색상, 텍스트 크기 등의 텍스트 속성은 데이터세트와는 관련이 없다고 판단했을 것으로 유추된다. 물론 NARA에서도 외관의 범주를 필수적인 영역으로 제시하지는 않았으나, 문자 인코딩과 텍스트 속성을 요소로는 제시하였다. 요약하면 데이터세트 유형의 필수보존속성 범주를 구조, 기능, 내용, 맥락의 범주에서만 고려하였다.

둘째, 기능 범주에서 '상호작용성'이라는 상위 개념을 제안하였다. 사용자는 SELECT, JOIN, CREATE, INSERT 등과 같은 SQL문을 통해 질의와 응답이라는 상호작용을 하게 된다. NARA에서는 조작 기능과 질의 표시를 필수속성으로 제시하는 반면, 국가기록원 R&D 연구결과에서는 좀 더 포괄적 개념으로 '상호작용성'을 제시하였다.

셋째, 내용 범주의 필수보존속성으로 복잡성(Complexity)과 이질성(Heterogeneity)을 제안하였다. 데이터베이스는 그 규모가 클수록 기능이 다양해지므로 관련한 데이터세트 요소가 증가하고 복잡해진다. 예를 들어 사용자 계정관리 기능, 권한 설정 기능 그리고 프로시져(Procedure), 함수(Function), 트리거(Trigger)와 같은 프로그래밍 요소들이 대표적인 예라

할 수 있다. 이렇듯 데이터뿐만 아니라 프로시저 등의 복잡성을 필수보존속성으로 제시하였다. 데이터베이스에 저장되는 디지털 객체는 정수형(INT, SHORT), 실수형(FLOAT, DOUBLE), 바이너리형(BLOB), 시간형(DATE, TIME) 등과 같은 정형 데이터뿐만 아니라 전자문서를 구성하는 문자형(CHAR, VARCHAR), 문장형(STRING, CLOB) 이외에도 이미지 파일과 같은 비정형 데이터를 포함하는 여러 가지 유형의 데이터 타입이 존재한다. 이와 같은 데이터베이스의 속성을 이질성으로 정의하였다.

넷째, 맥락 범주로는 행정정보 데이터세트 관리기준표를 필수보존속성으로 제안하였다. '행정정보 데이터세트 관리기준표'란 공공기관의 행정정보시스템을 통해 생산되는 데이터를 관리하기 위해 관련한 정보를 작성하도록 정한 서식을 말한다. 여기에는 정보시스템에 대한 관리정보, 시스템 및 생산 데이터와 관련한 법규정보, 행정정보시스템의 개요를 설명하는 시스템정보, 데이터베이스시스템의 대표 데이터 및 연계시스템의 유무 등을 파악할 수 있는 데이터의 정보, 어떤 업무 기반에서 생산된 데이터인지를 파악할 수 있는 업무정보, 데이터세트의 보존기간 및 기록관리와 관련한 기록관리정보가 있다.

다섯째, 구조 범주의 필수보존속성으로 관계성(Relationship)과 다양성(Diversity)을 제안하였다. NARA에서 제시한 데이터베이스 스키마(Database Schema), 기술 메타데이터(Technical Metadata)와는 다른 용어를 제시하고 있으나 그 개념은 비슷하다고 할 수 있다. 관계형 데이터베이스는 기본적으로 Table(Column, Row)로 구성되며, 테이블 간 관계(Relationship)가 기본키(PK)와 외래키(FK) 형태로 존재한다. 다수의 Table은 하나의 스키마 또는 데이터베이스에 포함되기도 한다. 이러한 테이블의 구성과 테이블 사이의 관계를 관계성이라고 정의하였으며, 이러한 구조의 보존을 필수보존속성으로 제시하였다. 보통 상용화된 데이터베이스들은 이러한 구조가 표준화되어 있지 않으며, 각자 다른 설계를 통해 구현되고 있다. 데이터베이스는 수시로 업데이트되기 때문에 여러 버전이 존재하는 특징을 가지는데, 예를 들어 Oracle에는 v5, v6…10g, 11g, 12c 등이 존재하며, MySQL 역시 1에서 8까지의 버전이 존재한다. 따라서 표준화되지 않은 다양한 구조와 여러 버전의 존재를 수용하는 개념으로 다양성을 정의하였다. 〈표 3〉은 국가기록원 R&D 연구결과와 NARA가 제안하는 데이터베이스 유형 전자기록의 필수보존속성을 비교한 것이다.

4. 데이터세트 유형 전자기록의 필수보존속성 도출

본 장에서는 3장의 내용을 바탕으로 데이터세트 유형 전자기록의 필수보존속성을 도출하고자 한다. 필수보존속성의 범주는 선행연구로 살펴본 전한역 외(2023) 논문에서 제시된 재현(Rendering), 기능(Behavior), 맥락(Context), 구조(Structure)를 범주로 한다(〈표 4〉 참조).

4.1 데이터베이스 유형 필수보존속성 도출

데이터베이스 유형 전자기록의 필수보존속성으로 9개의 요소를 제안한다.

〈표 3〉 국가기록원 R&D 연구결과와 NARA의 데이터베이스 유형 필수보존속성 비교

범주	NAK R&D	NARA
외관(Appearance)	-	• 문자 인코딩(Character Encoding) • 텍스트 속성(Text Properties)
기능(Behavior)	• 상호작용성(Interactivity)	• 조작 기능(Manipulation Functionality) • 질의 표시(Display Query)
내용(Content)	• 복잡성(Complexity) • 이질성(Heterogeneity)	-
맥락(Context)	• 행정정보 데이터세트 관리기준표	• 시리즈(Series) • 설명 메타데이터(Descriptive Metadata)
구조(Structure)	• 관계성(Relationship) • 다양성(Diversity)	• 데이터베이스 스키마(Database Schema) • 기술 메타데이터(Technical Metadata)

〈표 4〉 필수보존속성 관련 범주의 재정의

범주	의미
Rendering	• 데이터베이스(구조화데이터)가 출력 장치를 통해 볼 수 있는 데이터베이스(구조화데이터) 형태로 재현되는데 필요한 속성
Behavior	• 데이터베이스(구조화데이터)에 내장되어 이용자 혹은 이용개체 사이에서 이루어지는 상호작용할 수 있는 기능
Context	• 데이터베이스(구조화데이터)의 내용이나 생산 환경을 설명하기 위한 설명(descriptive)·기술(technical)·관리(administrative) 메타데이터
Structure	• 데이터베이스(구조화데이터)를 구성하는 요소를 설명하는 속성

첫째, 랜더링의 범주에서는 문자 인코딩(Character Encoding)을 필수보존속성으로 제안한다. 문자 인코딩은 데이터베이스에 저장된 문자, 문자열, 정수, 실수, 날짜, 시간 등의 표현 정보를 나타내는 속성으로 EBCDIC, ASCII, EBCPAC, Binary, Zone Decimal 등이 있다. 디지털자료를 장기보존하기 위한 표준 중 하나인 OAIS 참조모형에서도 디지털 객체를 필요한 시점에서 다시 식별하고, 이해할 수 있으려면, 식별 및 이해 가능한 '무엇'을 계속해서 재생산할 수 있도록 해주는 또 다른 디지털 객체가 필요하다고 하였다(한국기록관리학회, 2018). 문자 인코딩은 데이터베이스를 재현하고 내용을 이해하기 위한 필수적인 속성이라 할 수 있다. NARA에서 제시된 문자 인코딩은 외관 범주에 속한 속성이었기 때문에 '기록물의 시각적 표현과 관련한 특성'이라는 범주의 의미가 적용되었었다. 이러한 이유로 필수적인 보존속성에서 제외되었다. 그러나 필수보존속성의 범주를 외관(Appearance)에서 랜더링(Rendering)으로 재정의하였을 때는 '정보 객체의 퍼포먼스를 재현하는 데 기여하는 모든 정보'로 범주의 의미가 확장된다. 다시 말하면, 정보를 재이용하는 관점에서 데이터베이스가 출력 장치를 통해 이용자가 볼 수 있는 형태로 재현되는데 필요한 속성으로 필수보존속성의 범주가 재정의

되었을 때 도출될 수 있는 속성인 것이다. 문자 인코딩을 보존포맷 선정을 위한 고유기준의 평가 요소로 선정한다면 해당 보존포맷이 다양한 언어의 문자, 숫자, 부호, 기호, 객체 등을 표현하는 데이터 인코딩 방식을 지원하는지에 대한 평가가 필요할 것이다. NARA에서는 문자 인코딩 이외에 폰트, 글자크기, 색상 등과 같은 텍스트 속성(Text Properties)도 제시되었는데, 데이터세트는 사람이 아닌 컴퓨터에 의해 처리되고 분석되므로 필수보존속성에서 제외하는 것이 합당하다고 판단된다.

둘째, 기능 범주에는 NARA에서 제시한 '질의 표시(Display Query)'를 그대로 수용하고 '질의 문장(Query String)'을 추가할 것을 제안한다. 기능 범주의 의미는 데이터베이스 또는 구조화데이터에 내장되어 이용자 혹은 이용개체 사이에서 이루어지는 상호작용을 말하는 것으로 국가기록원 R&D 보고서에서는 '상호작용성'으로 제시되고 있다. 즉, 질의 문장은 데이터베이스가 만들어진 본래의 목적을 확인할 수 있는 정보로 사용자 또는 사용자 인터페이스에서 데이터베이스에 원하는 답을 얻기 위해 작성된 문장을 의미한다. 관계형 데이터베이스의 질의 문장은 보통 SQL이고, NoSQL 데이터베이스의 경우는 프로그래밍에 사용된 API 또는 API를 포함하는 프로그래밍 코드까지도 질의 문장에 해당될 수 있다. 이때 질의 문장에 포함되어 있는, 테이블 관련 입력값으로 사용된 데이터는 데이터베이스에서의 정확한 위치를 확인하기 위해 부가적인 정보까지 포함하여야 할 것이다. 따라서 질의 문장을 필수보존속성으로 하는 고유기준의 평가는 해당 보존포맷이 예전 업무에서 활용했을 당시 사용했던 질의 문장을 실행할 수 있는 기능이나 그 특성을 지원할 수 있는지에 대한 여부일 것이다. 질의 표시는 데이터베이스가 만들어진 본래의 목적과 그 결과까지 확인할 수 있는 정보를 담고 있으므로 이용자들이 직접 접하면서 기록의 가치를 확인할 수 있는 속성이라 할 수 있다. 질의 표시의 유형은 보고서, 인증서, 공문서에서부터 그래프, 지도, 멀티미디어 그리고 사용자 인터페이스까지 다양한 형태로 존재할 수 있다. 일반적으로 데이터세트는 대용량・대규모로 존재하기 때문에 이관・보존・활용에 많은 인적자원과 물적자원이 투입된다. 따라서 기관의 자원 현황을 고려하여 질의 표시의 형태가 고려되어야 할 것이다. 질의 문장과 질의 결과 모두 각기 독립적인 가치와 내용을 담고 있으며, 질의 결과의 형식은 기록물의 가치와 사용자 인터페이스 기능의 중요도에 따라 단순한 문서 형태에서 실제 시스템에서 제공하는 그대로의 GUI로 재현하는 것까지 다양하게 고려될 수 있다. 따라서 질의 문장과 질의 표시는 별도의 필수보존속성으로 구분하여 정의할 필요가 있다. NARA에서 제시한 또 다른 필수보존속성에는 조작 기능이 있다. 조작 기능은 테이블 내, 테이블 간의 관계를 검사하는 기능으로 기록물에 포함된 기능이 아니라 기록물의 외부의 다른 응용프로그램에서 실행되는 기능으로 필수보존속성에는 해당하지 않을 것으로 판단된다.

셋째, 맥락 범주는 설명 메타데이터(Descriptive Metadata), 기술 메타데이터(Technical Metadata) 그리고 관리 메타데이터(Administrative Metadata)를 제안한다. 설명 메타데이터는 NARA에서도 제시된 맥락 범주의 필수보존속성으로 데이터베이스의 명칭, 목적, 주제어, 요약, 기관명,

부서명, 담당자 등을 포함하는 데이터 생산의 맥락이나 내용을 이해하는 데 필요한 정보를 말한다. 즉, 이용자가 데이터베이스에 대한 기술적인(technical) 배경지식 없이 데이터베이스의 내용을 이해하는 데 도움을 주는 정보로 향후 이용자가 연구, 업무 등에 활용하기 위한 목적으로 데이터베이스의 내용을 참조하고 검색하는 데 필수적인 속성이라고 할 수 있다. 기술 메타데이터는 데이터베이스 형식, 소프트웨어, 소프트웨어 버전 등을 설명하는 메타데이터를 말한다. 대부분 데이터베이스에서 내려받기(export) 기능을 통해 생성되는 파일의 헤더에 자동 포함되며, 파일 형태로 이관했을 경우, 데이터베이스로 재현하기 위해 필요한 속성을 말한다. NARA에서는 디지털정지이미지(Digital still images), 동영상/디지털 시네마(Moving Image/Digital Cinema), 동영상/디지털 비디오(Moving Image/Digital Video) 유형의 전자기록에서는 맥락의 범주에서 기술 메타데이터를 포함하고 있으며, 데이터베이스와 구조화데이터 유형에서는 기술 메타데이터를 구조 범주에 포함하고 있다. 그러나 맥락 범주가 전자기록의 내용을 이해하는 데 필요한 설명이나 해당 전자기록이 생산된 생산 시스템과 소프트웨어의 특성을 파악하는 데 도움을 주는 속성으로 구성된다는 점에서 데이터베이스 유형에서도 역시 구조 범주보다는 맥락의 범주에 해당하는 속성으로 제시되는 것이 타당하다. 관리 메타데이터는 식별자, 저장 위치, 공개여부, 접근방법, 보존기간, 보존기간 책정사유, 권한 등 기록으로서 관리하기 위한 요소로 구성된다. 국가기록원 R&D 연구결과에서는 행정정보 데이터세트 관리기준표를 필수보존속성으로 제시하였는데, 여기서 행정정보 데이터세트 관리기준표는 관리 메타데이터의 일종으로 간주할 수 있으므로 필수보존속성의 용어로 사용하기에는 다소 무리가 있어 보인다. NARA에서 제시한 또 하나의 속성인 시리즈(Series)는 관리 메타데이터의 항목에 해당된다.

넷째, 구조 범주는 데이터베이스를 구성하는 요소를 설명하는 속성으로 스키마(Schema), 루틴(Routine), 연결(Linkage)이 제시될 수 있을 것이다. 스키마는 데이터의 구조 또는 데이터베이스의 설계를 의미하는 용어로 데이터베이스 스키마는 데이터베이스에 데이터가 저장되는 구조를 말하는 것으로 데이터베이스를 재현할 때 데이터들 사이의 관계와 의미를 확인할 수 있는 중요한 속성이다. 여기에는 문자, 문자열, 정수, 실수 등의 데이터 형식과 PK, FK 등의 데이터들 사이의 관계 그리고 열 속성, 행 정보 등과 같은 데이터 그룹화 정보가 포함된다. 관계형 데이터베이스에서는 이러한 정보를 테이블 간의 관계를 설명해주는 다이어그램인 ERD 및 테이블 명세서 등에서 확인할 수 있으며, 사용자 접근 권한, 보안, 무결성 등에 관한 정의까지도 스키마에 포함된다. 스키마를 필수보존속성으로 고유기준 평가항목을 도출 시에는 해당 파일 포맷이 데이터베이스를 구성하는 요소의 키(key), 값, 속성 등을 표현하는 방식을 지원하는지, 데이터베이스를 구성하는 요소들을 계층적으로 표현하는 구조를 지원하는지, 구조들 사이의 관계를 표현하는 방식을 지원하는지, 사용자의 계정, 접근 권한, 보안, 무결성 등 관리 또는 제어에 관한 내용을 표현할 수 있는 방식을 지원하는지 등에 대한 평가가 이루어져야 할 것이다. 루틴은 데이터베이스의 데이터

에 대한 간단한 계산, 특정 작업 수행을 수행하는 속성으로 프로그래밍을 통해서 원하는 계산, 작업 등을 수행할 수 있는 기능(프로시저, 함수, 트리거 등)이므로 해당 코드를 필수보존속성으로 명시하고 보존하여야 한다. 연결은 데이터를 데이터베이스 내에서 직접 저장·관리되지 않고, 외부에 저장된 데이터 객체에 대한 경로를 데이터베이스에 저장하거나 다른 데이터베이스에 연계하고 있는 경우에 필요한 속성이다. 데이터세트를 이관할 때는 이들 데이터까지 함께 보존할 수 있어야 하므로 필수보존속성에 포함해야 할 것이다. 앞서 언급한 바와 같이 NARA에서는 기술 메타데이터를 구조 범주에서 제시하고 있었다. 국가기록원에서는 기록물과 메타데이터를 구분하여 관리하고 있으므로 일관된 방식으로 관리해야 한다는 측면에서 기술 메타데이터는 다른 유형의 전자기록물과 동일하게 맥락의 범주에 배치하는 것이 타당할 것으로 생각된다. 이상으로 데이터베이스 유형의 필수보존속성을 정리하면 〈표 5〉와 같다.

〈표 5〉 데이터베이스 유형의 필수보존속성(안)

범주	필수보존속성	정의 및 설명	
Rendering	문자 인코딩 (Character Encoding)	정의	• 데이터베이스에 저장된 문자, 문자열, 정수, 실수, 날짜, 시간 등의 표현정보를 나타내는 속성
		설명	• 데이터베이스에 저장된 문자, 문자열, 정수, 실수, 날짜, 시간, 객체 등의 표현된 방식(ASCII, UNICODE, EBCDIC, Binary 등)를 명시하고 지원할 수 있어야 함
Behavior	질의 문장 (Query String)	정의	• 사용자 또는 인터페이스에서 데이터베이스에 원하는 답을 얻기 위해서 작성되는 문장
		설명	• 사용자 또는 사용자 인터페이스에서 데이터베이스로부터 답을 얻기 위해서 작성된 질의 문장으로, 관계형 데이터베이스의 SQL(Structured Query Language)이 대표적임 • 질의 문장에 포함된 인수들에 대한 정보(위치, 형식 등)까지 포함할 수 있으며, SQL 문장의 실행까지도 고려될 수 있음
	질의 표시 (Display Query)	정의	• 데이터베이스가 만들어진 본래의 목적과 그 결과까지 확인할 수 있는 정보
		설명	• 질의 표시의 유형은 보고서, 인증서, 공문서에서부터 그래프, 지도, 멀티미디어 그리고 웹페이지까지 다양한 형태의 사용자 인터페이스로 존재할 수 있음
Context	설명 메타데이터 (Descriptive Metadata)	정의	• 이용자가 데이터베이스에 대한 기술적인 배경지식이 없어도 데이터베이스의 내용을 이해하는 데 도움을 주는 정보
		설명	• 데이터베이스의 명칭, 목적, 주제어, 요약, 기관명, 부서명, 담당자 등이 포함될 수 있으며, 향후 이용자가 연구, 업무 등에 활용하기 위한 목적으로 데이터베이스의 내용을 참조하고 검색할 수 있는 정보
	기술 메타데이터 (Technical Metadata)	정의	• 데이터베이스 형식, 소프트웨어, 소프트웨어 버전 등을 설명하는 정보
		설명	• 데이터베이스에서 내려받기 기능을 통해 생성되는 파일의 헤더에 자동으로 포함되는 정보로 파일 형태의 이관 시 데이터베이스로 재현하기 위해 필요한 정보
	관리 메타데이터 (Administrative Metadata)	정의	• 데이터베이스의 데이터를 기록으로 관리하기 위한 정보
		설명	• 식별자, 위치, 공개여부, 보존기간, 보존기간 책정사유, 권한 등 기록관리와 관련한 정보

범주	필수보존속성	정의 및 설명	
Structure	스키마 (Schema)	정의	• 데이터베이스에 데이터가 저장되는 구조를 재현하고 데이터들 사이의 관계와 의미를 확인하기 위한 속성
		설명	• 관계형 데이터베이스의 경우 테이블의 정보, 테이블의 컬럼 정보(명칭, 형식), 데이터들 간 관계 정보(PK, FK 등)가 스키마에 해당되며, 사용자 접근 권한, 보안, 무결성 등에 관한 내용까지 스키마에 포함될 수 있음
	루틴 (Routine)	정의	• 데이터베이스의 데이터에 대한 간단한 계산, 특정 작업 수행과 관련한 속성
		설명	• 데이터베이스의 데이터에 대한 간단한 계산, 특정 작업 수행, 질의 문장 등을 수행할 수 있음. 관계형 데이터베이스의 경우, 저장 프로시저(Stored Procedure), 함수(Function), 트리거(Trigger) 가 대표적인 루틴임
	연결 (Linkage)	정의	• 외부와 연계된 데이터 객체(파일, 데이터베이스 등)와의 연결과 관련한 정보
		설명	• 데이터베이스 외부에 저장된 데이터 객체(파일, 페이지 등)에 대한 위치를 데이터베이스에 저장하거나 다른 데이터베이스에 연계하고 있는 상황에 대한 정보가 이에 해당되며, 데이터 객체 자체까지 포함될 수 있음

4.2 구조화데이터 유형의 필수보존속성

구조화데이터 유형의 필수보존속성은 총 7개가 도출되었다. 이 중 문자 인코딩, 설명 메타데이터, 기술 메타데이터, 관리 메타데이터, 스키마, 연결 등 6개의 필수보존속성은 앞에서 도출한 데이터베이스 유형의 필수보존속성과 그 정의와 설명이 중복된다. 기능 범주의 필수보존속성으로는 NARA에서 제시했던 하이퍼링크(Hyperlinks) 대신에 '검증(Verification)'을 제안한다. 문서유형 파일포맷에서는 하이퍼링크를 일반 텍스트와는 별도의 객체로 구분하고 있는데, 예를 들어 아래아 한글에서 URL작성 시 자동으로 파란색의 밑줄 서식이 설정되며, 삭제할 경우에는 하이퍼링크의 삭제 여부를 묻는 팝업창이 생성되는 것이 그것이다. 그러나 구조화데이터에서의 하이퍼링크는 별도의 객체가 아닌 텍스트로 표현되는 정보이므로 제외하는 것이 합리적일 것이다. 검증은 문법의 준수 여부, 특정 양식에 대한 유효성 등의 검증을 말한다. 데이테베이스 유형은 이미 데이터베이스의 무결성 제약조건을 기통과한 데이터세트가 변환 또는 이관되므로 데이터 자체가 잘 이관된다면 데이터 형식과 구조에 대한 별도의 검증은 고려될 필요가 없을 것이다. 그러나 구조화데이터는 검증여부를 보장할 수 없으므로 구조화데이터의 내용이 형식에 맞게 작성되었는지를 판단하는 작업은 해당 기록물을 이해하는 데 필수적이라고 할 수 있다. 데이터를 저장하거나 전송하기 위한 언어는 대부분 작성의 규칙과 문법이 존재한다. CSV는 데이터를 콤마(,)로 구분하여 저장하거나 JSON은 데이터를 속성-값 쌍의 집합으로 표현한다, XML은 HTML과 비슷한 태그 기반 구조로 되어 있으며 태그 안에 값을 저장한다. 다른 유형의 파일포맷들은 대부분 전용 앱이 존재하며 내용의 확인이 문법의 준수 여부와 직결된다. 반면 구조화데이터는 대부분 텍스트 형식으로 일반 텍스트 편집기나 뷰어로는 확인할 수 있으나 이는 문법의 준수와는 별개 문제이다. 따라서 검증은 해당 양식에 대한 유효성을 검증하는 기능으로 정의될 수 있을 것이다. 검증을 위해서는 파일포맷이 작성된 구문

에 대한 문법 준수 여부를 검증할 수 있는 방식을 지원하는지, 또는 파일포맷이 특정 양식을 정의하고 해당 양식에 대한 유효성을 검증하는 방식을 지원하는지 등에 대한 평가가 필요하다. NARA에서 제시되었던 열 개수, 행 개수 그리고 관련 파일은 필수보존속성으로 따로 제시하기 보다는 기술 또는 설명 메타데이터의 세부에 포함시키는 것을 제안한다. 이상으로 제안한 구조화데이터 유형 전자기록의 필수보존속성의 정의와 내용을 정리하면 〈표 6〉과 같다.

〈표 6〉 구조화데이터 유형의 필수보존속성(안)

범주	필수보존속성	정의 및 설명	
Rendering	문자 인코딩 (Character Encoding)	정의	• 구조화데이터에 저장된 문자, 문자열, 정수, 실수, 날짜, 시간 등의 표현정보를 나타내는 속성
		설명	• 구조화데이터에 저장된 문자, 문자열, 정수, 실수, 날짜, 시간 등의 나타내는 표현 정보(EBCDIC, ASCII, EBCPAC, Binary, Zone Decimal 등)
Behavior	**검증 (Verification)**	정의	• 작성문법의 준수여부, 특정 양식에 대한 유효성 등의 검증을 지원하는 속성
		설명	• CSV, JSON, XML, HTML 등은 대부분 작성 규칙, 문법 등이 존재함. 이를 통해서 문법의 준수 여부 • XML의 경우에는 특정 양식(DTD, XML Schema 등)을 정의하고, 해당 양식에 대한 유효성을 검증하는 기능을 표준에 포함하고 있음
Context	설명 메타데이터 (Descriptive Metadata)	정의	• 이용자가 구조화데이터에 대한 기술적인 배경지식이 없어도 구조화데이터의 내용을 이해하는데 도움을 주는 정보
		설명	• 구조화데이터의 명칭, 목적, 주제어, 요약, 기관명, 부서명, 담당자 등이 포함될 수 있으며, 향후 이용자가 연구, 업무 등에 활용하기 위한 목적으로 구조화데이터의 내용을 참조하고 검색할 수 있는 정보
	기술 메타데이터 (Technical Metadata)	정의	• 특정 데이터베이스 형식, 소프트웨어, 소프트웨어 버전 등을 설명하는 정보
		설명	• 대부분의 구조화데이트의 파일 헤더에 자동으로 포함되어 있는 정보로 파일 형태로 이관할 경우, 구조화데이터로 재현하기 위해 반드시 필요한 정보 • 구조화데이터의 전체 열 개수 및 행 개수는 해당 기록물의 포맷 변환시 완료 여부를 검증할 때 의미있는 정보이므로 설명 메타데이터에 포함함 • 데이베이스 형식, 소프트웨어 및 버전 등은 향후 재현을 위한 정보로 활용될 수 있으므로 기술 메타데이터에 포함함
	관리 메타데이터 (Administrative Metadata)	정의	• 구조화데이터를 기록으로 관리하기 위한 정보
		설명	• 식별자, 위치, 공개여부, 보존기간, 보존기간 책정사유, 권한 등 기록관리와 관련한 정보
Structure	스키마 (Schema)	정의	• 구조화데이터에 데이터가 저장되는 구조를 재현하고 데이터들 사이의 관계와 의미를 확인하기 위한 속성
		설명	• 각 파일포맷마다 제공하는 스키마의 구조가 다를 수 있음 • CSV의 경우 2차원 테이블구조를 제공하며, XML과 JSON은 키-값으로 구성된 요소를 트리형태의 계층적인 구조로 제공할 수 있음
	연결 (Linkage)	정의	• 외부와 연계된 데이터 객체(파일, 데이터베이스 등)와의 연결과 관련한 정보
		설명	• 데이터를 구조화데이터 내에서 직접 저장 관리하지 않고 외부에 저장된 데이터 객체에 경로를 만들어 데이터베이스에 저장하거나 다른 데이터 객체에 연계하는 경우, 데이터세트를 이관할 때 이들 데이터까지 함께 보존하기 위한 정보

5. 결 론

본 연구는 현재까지의 전자기록 장기보존에 대한 정책이 문서유형 위주의 전자기록물에 치중되어 온 점과 다양한 행정정보시스템을 통해 생산되는 문서유형 이외의 전자기록 장기보존에 대한 문제 인식에서 시작되었다. 특히, 데이터 관리에 관한 사회적 관심이 증대되는 시점에서 데이터세트에 대한 장기보존포맷 선정체계에 대한 연구가 필요한 시점이라고 할 수 있다. 전자기록의 보존포맷선정을 위한 고유기준은 이미 마련되어 있으므로 전자기록의 유형 중의 하나인 데이터세트의 고유기준이 마련되어야 한다. 이를 위해 해외의 사례로는 유일하게 데이터세트 유형 필수보존속성을 제시한 미국의 NARA와 국가기록원 R&D 연구보고서를 중심으로 연구를 진행하였다. 다만, 다양한 국가의 사례를 비교하지 못했다는 점은 연구의 한계로 남는다. 연구의 결과로 데이터세트를 데이터베이스 유형과 구조화데이터 유형으로 구분하고 데이터베이스형 필수보존속성으로 9개의 요소와 구조화데이터형 필수보존속성으로 7개의 요소를 도출하였다.

전자기록물의 장기보존은 오랜 시간이 경과하거나 해당 기록을 재현하는 기술적 환경이 달라지더라도 재현되어야 하는 것이 핵심일 것이다. 따라서 기록관리기관은 재현의 시점에서 언제든지 기록에 접근할 수 있는 하드웨어 및 소프트웨어적 환경을 마련하여야 한다. 또한 전자기록의 유형에 따른 디지털 객체에 관한 기술적 특징을 파악하고 있어야 할 것이다. 본 연구의 결과가 데이터세트 유형의 보존포맷 선정을 위한 고유기준 체계를 마련하는 데 기초 자료로 활용되기를 바란다.

참 고 문 헌

공공기록물의 관리에 관한 법률 시행령. 대통령령 제33575호.

공공기록물의 관리에 관한 법률. 법률 제19408호.

국가기록원 (2019). 데이터세트 유형 전자기록의 장기보존기술 연구.

국가기록원 (2020). 문서유형 보존포맷 및 장기보존패키지 다양화 연구.

국가기록원 (2022a). 전자기록물 보존포맷 선정기준(v.1.0).

김명옥, 리상용 (2010). 전자기록물의 장기보존을 위한 기능요소 연구. 한국기록관리학회지, 10(2), 101-126.

나시다 케이스케 (2018). 빅데이터를 지탱하는 기술. 경기: 제이펍

남성운, 윤대현 (2001). 전자기록물의 장기보존을 위한 방안 연구: 개념을 중심으로. 한국기록관리학회지, 1(2), 101-120.

박우창, 남송휘, 이현룡 (2019). MySQL로 배우는 데이터베이스 개론과 실습. 서울: 한빛아카데미.

소정의, 한희정, 양동민 (2018). 국외 전자기록물의 장기보존 정책 비교 분석: 미국, 캐나다, 영국, 호주, 스위스를 중심으로. 한국기록관리학회지, 18(4), 125-148. http://doi.org/10.14404/JKSARM.2018.18.4.125

오세종 (2023). DB 설계 입문자를 위한 데이터베이스 설계 및 구축. 서울: 생능출판사.

전한역, 김지혜, 김현태, 양동민 (2023). 시청각 유형 보존포맷 선정을 위한 필수보존속성 연구: 디지털 오디오를 중심으로. 디지털문화아카이브지, 6(2), 27-53. http://doi.org/10.23089/jdca.2023.6.2.002

한국기록관리학회 (2018). 기록관리의 이론과 실제. 서울: 조은글터.

한국정보통신기술협회 [발행년불명]. 정보통신용어사전. 출처: http://word.tta.or.kr/dictionary/searchList.do

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

Elmasr, R. & Navathe, S. (2015). Fundamentals of Database Systems (7th ed.). Boston: Pearson.

InSPECT (2009. October 13). InSPECT Framework Report. Available: https://significantproperties.kdl.kcl.ac.uk/methodology.html

NAA (2022). Preservation Digitisation Standards. Available: https://www.naa.gov.au/about-us/our-organisation/accountability-and-reporting/archival-policy-and-planning/preservation-digitisation-standards#preservation-digitisation-standards.

NARA (2009. October 26). Significant Properties. Available: https://www.archives.gov/files/era/acera/pdf/significant-properties.pdf

NARA (2022a. July 14). Preservation Action Plan for Structured Data/Databases National Archives and Records Administration. Available: https://github.com/usnationalarchives/digital-preservation/blob/master/Structured_Data_Formats/NARA_PreservationActionPlan_Databases_20220714.pdf

NARA (2022b. July 14). Preservation Action Plan: Structured Data National Archives and Records Administration. Available: https://github.com/usnationalarchives/digital-preservation/blob/master/Structured_Data_Formats/NARA_PreservationActionPlan_StructuredData_20220714.pdf.

NARA (2022c. July 14). Preservation Action Plan for Structured Data/Calendars National Archives and Records Administration. Available: https://github.com/usnationalarchives/digital-preservation/blob/master/Structured_Data_Formats/NARA_PreservationActionPlan_Calendars_20220714.pdf

NARA (2022d. July 14). Preservation Action Plan: Structured Data/Spreadsheets National Archives and Records Administration. Available: https://github.com/usnationalarchives/digital-preservation/blob/master/Structured_Data_Formats/NARA_PreservationActionPlan_Spreadsheets_20220714.pdf.

TNA (2017). Digital Strategy 2017-2019.

TNA (2023, August 3). Archives for Everyone 2023-27 published. Available: https://www.nationalarchives.gov.uk/about/news/archives-for-everyone-2023-27-published/

• 국문 참고자료의 영어 표기
(English translation / romanization of references originally written in Korean)

Enforcement Decree of the Public Records Management Act. Presidential Decree No.33575.

Han, Hee-jeong, Oh, Hyo-jeong, & Yang, Dongmin (2020). A study on the selection of preservation format for long-term preservation of electronic records. Journal of the Korean Society of Records Management, 20(1), 69-87. http://doi.org/10.14404/JKSARM.2020.20.1.069

Jeon, Hanyeok, Kim, Ji-Hye, Kim, Hyun-Tae, & Yang, Dongmin (2023). A study on significant properties for selection of audiovisual type preservation format focused on digital audio. Journal of Digital Cultural Archives, 6(2), 27-53. http://doi.org/10.23089/jdca.2023.6.2.002

Kim, Myeong-Ok & Lee, Sang-Yong (2010). A study on the functional elements for long-term preservation of electronic records. Journal of Korean Society of Archives and Records Management, 10(2), 101-126.

Korea Telecommunication Technology Association [n.d.]. Dictionary of Information and Communication Terms. Available: https://terms.tta.or.kr/dictionary/dictionaryView.do?subject=%EB%8D%B0%EC%9D%B4%ED%84%B0+%EC%84%B8%ED%8A%B8

Korean Society of Records Management (2018). Records and aArchives Mnagement: Theory and practice. Seoul: Joeun glitter.

Nam, Sung-Un & Yoon, Dai-Hyun (2001). A study of the methodology for the long-term preservation of electronic records: focus on the preservation concept. Journal of Korean Society of Archives and Records Management, 1(2), 101-120.

Nasida Keisuke (2018). Technology that supports big data. Gyeonggi: J Pub.

National Archives of Korea (2019). Study on long-term preservation technology of dataset-type

electronic records.

National Archives of Korea (2020). Study on Diversification of the Document-Type Preservation Format and Long-term Preservation Package.

National Archives of Korea (2022a). Selection Criteria for Preservation Format of Digital Records (Version 1.0).

Oh, Sejong (2023). Database Design and Construction for DB Design Beginners. Seoul: Life & Power Press Co.

Park, Woo-chang, Nam, Songhwi, & Lee, Hyeonryong (2019). Database Introduction and Practice with MySQL. Seoul: Hanbit Media.

Public Records Management Act. No. 19408

So, Jeong-Eui, Han, Hee-jeong, & Yang, Dongmin (2018). A comparative analysis of long-term preservation policies in foreign electronic records: nara, lac, tna, naa, and sfa. Journal of the Korean Society of Records Management, 18(4), 125-148. http://doi.org/10.14404/JKSARM.2018.18.4.125