# 행정정보 데이터세트 이관도구 SIARD_KR의 개선방안*

## Improvement of Administration Information Dataset Transfer Tools 'SIARD_KR'

변우영 (Woo-Yeong Byeon)**
임진희 (Jin-Hee Yim)***

### 초 록

SIARD_KR은 스위스 연방 기록보존소에서 개발한 관계형 데이터베이스 컨텐츠의 장기보존에 이용하는 기술인 SIARD를 우리나라의 실정에 맞게 일부 수정한 행정정보 데이터세트 보존 도구이다. 기존의 선행연구는 SIARD가 얼마나 관계형 데이터베이스안에 들어있는 모든 데이터를 손실 없이 잘 추출할 수 있는지에 초점이 맞춰져 있다. 하지만 데이터베이스에 들어있는 데이터 전부가 의미 있는 정보, 즉 행정정보 데이터세트는 아니다. 따라서 이 논문은 SIARD_KR이 행정정보 데이터세트의 특성을 반영하고 있는가에 대한 문제의식에서 시작한다. SIARD_KR이 단순히 DB에 저장된 데이터를 추출하는 도구가 아니고 의미 있는 정보만을 식별하여 추출할 수 있을지, 본래의 시스템에서 유리되어도 의미 있는 정보를 유지할 수 있을지 확인하려 한다. 본 논문은 SIARD_KR의 구조를 분석하고, 예상되는 문제점을 도출하여 그에 대한 개선방안을 제시하는 것을 목적으로 한다.

### ABSTRACT

SIARD_KR is an administrative information dataset preservation tool. It is a partially modified version of SIARD, technology used for long-term preservation of relational databases developed by the Swiss Federal Archives, to suit Korea's situation better. Previous studies have focused on how SIARD is able to effectively extract all data contained in the relational database without loss. However, not all data contained in the database is meaningful information, that is, an administrative information dataset. This paper began, therefore, with the awareness of the problem of whether SIARD_KR reflects the characteristics of the administrative information dataset. SIARD_KR is not only a tool for extracting data stored in the DB. We want to see if it is capable of identifying and extracting only meaningful information, and maintaining meaningful information, even if it is separated from the original system. The purpose of this paper is to analyze the structure of SIARD_KR, identify expected problems, and suggest improvement measures for them.

키워드: 행정정보 데이터세트, SIARD, SIARD_KR, 이관도구, 데이터세트, 데이터베이스
administration information dataset, SIARD_KR, siard, transfer tools, dataset, database

---

* 본 연구는 대학원 석사학위논문을 수정・요약한 것임.

** 명지대학교 기록정보관리학과 석사과정(WMBRyeon@gmail.com) (제1저자)
*** 명지대학교 기록정보과학전문대학원 조교수(yimjhkr@mju.ac.kr) (교신저자)

■ 논문접수일자 : 2022년 2월 14일 ■ 최초심사일자 : 2022년 3월 2일 ■ 게재확정일자 : 2022년 3월 10일
■ 정보관리학회지, 39(1), 195-217, 2022. http://dx.doi.org/10.3743/KOSIM.2022.39.1.195

# 1. 서 론

행정정보 데이터세트는 국가기록원(2020a)이 공공기관의 행정정보시스템에서 생산, 수집, 가공, 저장, 검색, 제공, 송신, 수신 등을 위해 조합된 문자, 숫자, 도형, 이미지 및 그 밖에 데이터 집합이라고 정의하고 있다. 정의가 나타내는 바와 같이 행정정보 데이터세트는 일반적인 문서유형의 기록과는 다른 특성을 지닌다. 심지어 업무관리시스템에서 생산되는 전자문서라도 기존의 종이기록을 전자적 환경에서 처리하려는 목적을 가지고 설계되었기 때문에(임진희, 2021) 형태만 전자적일 뿐, 특성은 종이문서에 가깝다.

종래의 기록과 달리 행정정보 데이터세트는 다음과 같은 특성을 추가적으로 갖는다. 첫 번째로, 시스템 종속적이라는 특성은 행정정보시스템에서 데이터세트가 유리되었을 때, 원래의 룩앤필을 보존하기 힘들다. 기존의 문서유형의 기록에서 서식에 해당하는 부분을 바로 행정정보시스템에서 구현해주기 때문이다. 또한 유리된 데이터세트를 다른 행정정보시스템에 탑재한다고 해서 원래의 기능을 갖게된다는 보장 또한 없다. 따라서 행정정보 데이터세트를 이관할 때에는 룩앤필에 대한 추가적인 고려가 필요하다.

두 번째로 모든 행정정보 데이터세트가 의미있는 정보가 아니라는 특성을 지닌다. 이것은 모든 행정정보 데이터세트가 보존의 대상이 아니라는 말과 같다. 문서유형의 기록은 구성요소인 서식과 내용 모두가 업무를 표상한다. 하지만 행정정보 데이터세트의 경우 어떠한 업무를 어떤 행정정보 데이터세트가 표상하는지에 대한 식별과정이 필요하다. 식별과정을 거쳐 이름을 얻은 행정정보 데이터세트는 업무를 표상하는 의미 있는 정보가 될 수 있지만, 그중에서도 보존의 가치가 있는 행정정보 데이터세트는 일부이다. 따라서 반드시 DB안의 모든 데이터가 보존의 대상이 될 필요는 없다.

마지막으로 행정정보 데이터세트는 동적인 상태로 존재한다는 특성이다. 폐지된 기관이 아니라면, 행정정보시스템은 기관의 업무를 수행하기 위해서 지속적으로 사용된다. 그러므로 행정정보 데이터세트 또한 지속적으로 변화하는 상태이다. 따라서 이관을 주기적으로 시행해야 하는 경우 중복파일에 대한 문제를 고려해야 한다.

앞서 언급한 행정정보 데이터세트의 특성으로 인하여 2010년에 공공기록물 관리에 관한 법률(이하 공공기록물법)이 개정되면서 행정정보 데이터세트 기록이 기록관리의 범위로 포착되었으나, 실질적인 관리는 이루어지지 못했다. 특히 기존의 기록관리 체계는 종이기록으로 대표되는 종래의 재래식 기록에 치중하여 설계되었다는 점이 그 원인일 것이다. 국가기록원은 이 상황을 해결하고자 2016년부터 현재에 이르기까지 행정정보 데이터세트의 기록관리 방안을 설계하려는 연구개발을 지속적으로 수행하였으며, 그 결과로 2020년 3월 공공기록물법 시행령에 행정정보 데이터세트 관리와 관련된 조항이 신설되면서 'NAK 35 : 2020 행정정보 데이터세트 기록관리기준 - 관리기준표의 작성 및 이관규격'이라는 국가기록원 표준이 제정되었다. NAK 35에서는 행정정보 데이터세트의 관리기준표 작성방법과 함께 이관규격으로 스위스 연방기록원의 SIARD를 제시하고, 이를 위한 도

구인 SIARD_KR을 이용할 것을 권장한다(국가기록원, 2020).

이 논문은 SIARD_KR이 행정정보 데이터세트의 특성을 반영하고 있는가에 대한 문제의식에서 시작한다. SIARD_KR이 단순히 DB에 저장된 데이터를 추출하는 도구가 아니라 의미있는 정보만을 식별하여 추출할 수 있을지, 본래의 시스템에서 유리되어도 의미 있는 정보를 유지할 수 있을지, 나아가 본래의 시스템의 기능 또한 구현할 수 있는지 확인하려 한다. 따라서 본 연구는 SIARD_KR의 구조를 분석하고, 3장의 서두에서 후술할 샘플 DB를 이용하여 구조적 한계를 밝히고 개선방안을 모색하는 것을 목적으로 한다.

본 논문에서는 이관도구로 권장되는 SIARD_KR을 소개하고 한계점을 지적하기 위하여 먼저 행정정보 데이터세트의 관리방안과 관련된 선행연구를 통해서 행정정보데이터세트를 관리하기 위하여 추가로 고려되어야 하는 사항을 도출한다. 다음으로 SIARD와 관련된 논문을 통해서 SIARD의 특징과 한계점을 밝힌다. 이를 바탕으로 현행 이관도구인 SIARD_KR의 한계점을 지적하고 행정정보 데이터세트의 이관도구가 갖춰야할 추가적인 기능요건을 제시한다.

## 2. 선행연구

### 2.1 행정정보 데이터세트 관리방안 연구

조은희, 임진희(2009)는 행정정보시스템이 계속 증가하는 상황에서 행정정보 데이터세트가 기록관리의 사각지대에 놓여있음을 지적하고 전자정부 추진 전략의 틀에서 데이터세트 식별체계의 수립과 기록관리 기준의 설정 전략을 제안하고 있다. 데이터세트의 유형을 각종 통계 및 설문을 수행한 원자료(Raw Data), 카드 및 대장류, 전자문서와 업무트랜잭션 데이터가 복합된 유형, 관측데이터 유형의 4가지로 분류하고, 유형별로 서비스 특성을 제시하고 있다. 또한 행정정보 데이터세트를 기록으로 관리하기 위해서는 가치선별 이전에 식별의 과정이 선행되어야 한다고 한다. 행정정보 데이터세트의 식별이란 통제와 관리단위를 확정하고 계층구조를 구성하는 것으로 정의하고 있다. 이때 데이터베이스 자체나, 혹은 데이터베이스의 계층구조를 그대로 행정정보 데이터세트로 식별하는 것은 업무를 표상하는 기록의 요건을 만족시키기 어려우므로 부적절함을 밝히고 있다.

왕호성, 설문원(2017)은 데이터기록의 진본성을 위하여 기록의 내용, 구조, 맥락 뿐만 아니라 외형과 기능을 함께 보존하여야 한다고 한다. 이를 위해서 재현성을 강조하고 이를 실현하는 도구로 에뮬레이션 전략을 꼽았다. 하지만 데이터세트 기록관리의 최소단위가 다양해지면 실제 이루어지는 기록관리 행위의 설계가 어려워진다고 주장하였다. 또한 데이터베이스는 이미 고도로 구조화되어 있으므로 데이터베이스의 구조 외의 최소단위를 정의하는 것은 위험부담이 크다고 주장하였다.

오세라, 박승훈, 임진희(2018)는 행정정보 데이터세트 기록관리 필요성은 기록관리 연구자들 사이에서 넓은 공감대를 형성하고 지속적으로 연구되어 왔지만 실행의 단계에 이르지 못한 것은 현실 적용이 가능한 관리방안의 부재에 있음을 지적하고 당시의 행정정보시스템에

서의 데이터세트 생산 및 관리 환경 사례를 조사하여 행정정보 데이터세트 관리방안을 개발하는데 밑거름이 될 수 있는 조사방법론을 제시하고 있다. 해당 연구에서는 국가기록원에서 제시하는 사용기관 사이 관계에 따라 분류한 행정정보시스템 유형을 이용하여 1~2개의 시스템을 선정하고, 현장조사와 소프트개발방법론 중 하나인 점진적 나선형 분석방법을 적용하여 분석을 진행하였다. 분석의 결과로 데이터세트 기록관리 방안은 현행 전자문서와는 다르게 접근해야 하며, 데이터세트는 단순히 행과 열의 구조가 아닌 수많은 포맷의 전자파일도 포함되어 있음을 도출하였다.

오세라, 이해영(2019)은 데이터세트의 경우 기술 종속적인 특성과 방대한 규모 및 시스템별 다양한 운영환경으로 인하여 현행 종이기록 중심의 표준 기록관리 지침과 절차를 적용하기 어려운 점을 지적하고, 업무현장에서 적용할 수 있는 관리방안과 절차를 개발하였다. 이 연구에서는 데이터세트 기록관리가 방치된 가장 큰 원인을 태생이 문서류와는 다른 데이터세트에 문서류와 같은 기준과 관리방법을 적용하려는 데에 있다고 보았다. 특히 데이터베이스 분석에서 물리적 데이터의 형식을 분석하여 데이터가 해당 시스템에서 분리될 경우 해석하기 어려운 형식이 존재하는지를 체크하고 물리적 테이블의 관계 분석을 통해 데이터 그룹 및 엔티티의 최소단위를 도출하였다.

노명환(2020)은 행정정보 데이터세트 기록의 보존을 위해서 반드시 이관이 동반되어야 하는 것은 아니라고 한다. 이관이라는 행위는 맥락이 흐트러져 데이터 값이 변하는 상황을 초래할 수 있으며, 예외적인 경우에만 시행되어야 한다고 한다. 바로 그 예외적인 경우에 SIARD_KR을 적용하는 것은 신뢰할 수 있지만, 그렇다 하더라도 이관시에 맥락의 변화는 피할 수 없으며 원래의 상태와는 손색이 있을 것이라고 한다.

류한조, 백영미, 임진희(2021)는 국가철도공단의 재산관리시스템을 중심으로 데이터세트 생산시스템의 기능요건을 연구하였다. 행정정보시스템은 객체의 단위를 분명하게 구분하여 식별할 수 있는 문서기반시스템하고는 다르므로 기존의 기록관리 원칙을 적용하기 어렵다고 한다. 또한 생산시스템에서 장기간 데이터세트를 보존하고 그 과정에서 처분 등 일부 기록관리 업무가 발생하기 때문에 기록관리 기능요건 및 업무시스템 요건을 모두 고려해야 한다고 한다. 따라서 기록생산 시스템과 기록관리시스템에 대한 기능요건과 데이터세트의 특성을 고려하여 진본성・무결성・신뢰성, 데이터품질, 기록관의 3가지 측면에서 기능요건을 제시하였다. 이 과정에서 데이터세트와 테이블의 관계가 多 대 多 관계를 맺거나 데이터세트의 종결시점을 명확히 정의하기 어렵다는 등의 특징을 도출하였다.

황진현, 백영미, 임진희(2021)는 국가철도공단의 전자조달 시스템을 중심으로 행정정보시스템 내 데이터세트 기록의 식별 및 관리기준표 작성, 평가, 보존기간 책정 및 (영구)기록관리시스템으로의 이관여부 결정까지의 과정 중 필요한 절차 및 도구 등을 개발하고 설명하였다. 해당 논문에서는 조은희, 임진희(2009)의 연구와 동일하게 데이터세트 기록을 관리하기 위해선 가장 먼저 가치평가 이전에 데이터세트의 어느 범주를 하나의 기록으로 볼 것인지 식별하고 확정해야 한다고 한다. 이 과정에서 시

스템 내 모든 데이터들을 데이터세트 기록으로 식별하고 관리방안을 마련해야 함을 강조하였다. 또한 종래의 보존기간이 만료된 시점에 폐기를 위한 평가를 하는 기록과 달리 데이터세트의 경우 ISO 15489-1:2016의 확장된 평가의 정의를 수용하여 사전평가가 이루어져야 한다고 한다.

이경남, 최광훈, 임진희(2021)는 국가철도공단의 사업관리 시스템을 중심으로 데이터세트 기록을 식별하고 평가한 과정과 결과를 분석하여 데이터세트 기록관리 프로세스를 실증적으로 검증하고 이를 위한 기록관의 역할을 연구하였다. 이 과정에서 동일한 데이터세트가 여러 시스템에 존재하는 것을 확인하고 원천시스템의 중요성을 강조하였다. 뿐만 아니라 시스템 간에 데이터가 연계되어 있을 경우 연결된 모든 시스템을 확인하고, 데이터를 비교하여 고유한 데이터를 식별하고 필요시 연계시스템의 데이터세트를 하나의 데이터세트로 식별하는 것도 필요함을 밝혔다.

송치호, 임진희(2022)는 데이터세트 기록을 구성하는 원자료(Raw Data)의 품질이 보장되지 않으면 데이터세트, 더 나아가 기록자체를 신뢰할 수 없음을 바탕으로 현재 범정부 적으로 시행되고 있는 공공데이터영역의 품질관리 체계와 비교하여 기록관리관점에서의 평가영역과 지표, 평가 도구를 제시하고 이를 실제로 적용하였다. 특히 데이터의 품질평가의 목적을 데이터세트 식별 이후 내용적가치 및 데이터 연계가치에 대한 가치평가가 이루어진 데이터세트에 대해서 최종적으로 품질을 확인하도록 제시하였다. 이러한 방향성은 데이터베이스 전체에 대한 품질평가가 아닌 업무적으로 연계성이 있는 데이터만을 평가 대상으로 지정하여 모든 데이터가 행정정보 데이터세트로 관리될 필요성은 없음을 보여주었다.

### 2.2 스위스 연방 기록보존소의 SIARD에 관한 연구

SIARD(Software Independent Archival of Relational Databases)는 관계형 데이터베이스에 저장되어 있는 데이터세트를 소프트웨어와 독립적으로 하나의 파일로 '장기보존' 할 수 있도록 개발된 공개표준이다. 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 2.2까지 업데이트 되어있으며, 본 논문에서는 SIARD Suite 및 SIARD_KR의 바탕이 된 2.1 버전을 기준으로 한다. SIARD는 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 가능하다는 특징이 있다(한희정 외, 2020).

RDB와 같은 DB에 저장된 정보를 시스템이나 DB 외부에 저장하는 방법으로 덤프라는 방법은 존재해왔다. 하지만 원래의 데이터베이스를 복원해야 할 필요성이 있다면, 덤프된 데이터를 가지고 원래의 데이터베이스 구조를 리버스엔지니어링(reverse engineering)하여 다시 데이터베이스를 설계하고 덤프된 데이터를 입력해야 하는 불편함과 동시에 이전의 데이터베

이스와 동일한 형태를 반드시 취할 것이라는 보장은 없었다. 기존의 데이터베이스를 설계할 때 생산된 기록이 있다면 원래의 데이터베이스의 형태로 복원될 가능성이 높아지겠지만, 원래의 상태를 담보하기에는 부족했다. 뿐만 아니라 덤프 방식으로 데이터베이스를 스냅샷의 형태로 저장하면 이러한 관계형 데이터베이스의 특성을 나타내는 데이터베이스 스키마 정보들은 소실된 채, 데이터베이스 내부의 저장된 정보만을 저장한다.

관계형 데이터베이스(RDB: Relational DataBase)는 1970년대 IBM의 에드거 F. 커드(E. F. Codd)가 제안한 관계형 모델(relational model)을 바탕으로 개발된 데이터베이스의 형태이다. 관계형 모델은 실제 세계의 데이터를 수학적 논리 관계 개념을 사용하여 행(row)과 열(column)로 표현한 표(table)와 행과 열의 상관관계로 정의하는 데이터 모델이다. 이때, 단순히 표의 형태로 데이터를 표현하는 것에 그치지 않고, 기본키(Primary key, PK) 및 외래키(Foreign key, FK) 등의 개념과 제약조건, 데이터형식을 통해서 데이터베이스의 관계와 무결성을 지켜준다. 이렇게 데이터베이스를 구조화시켜주는 기본키나 외래키, 제약조건, 데이터형식 등이 스키마 정보이기 때문에 RDB에서 스키마 정보가 소실되는 것은 더이상 구조화된 의미 있는 정보가 아닌 데이터의 집합으로 변화함을 의미한다.

SIARD가 보존포맷이 될 수 있는 가장 큰 이유는 바로 단순히 데이터베이스의 정보를 덤프하여 저장하는 것이 아니라, 관계형 데이터베이스의 특징을 보존하고, 필요시 원래의 데이터베이스를 복원할 수 있다는 점이다. 또한 대부분의 행정정보 데이터세트가 RDB 형태로 저장되고 있다는 점에 착안하여 SIARD를 행정정보 데이터세트의 보존포맷으로 사용하기 위한 적합성을 연구하는 선행연구가 있었다.

이 영역에서는 스위스 연방 기록보존소에서 개발한 SIARD의 개발 배경, 원리, 기능, 활용사례 등을 다루고 있다. 김주연(2020)은 스위스 연방 기록보존소에서 개발한 관계형 데이터베이스 컨텐츠 장기보존 기술인 SIARD를 연구하고, 국내 행정정보 데이터세트 기록관리 및 장기보존에 적용하는 방안을 모색한다. 이를 위해서 SIARD의 주요기능과 활용사례를 살펴보고, 국내에 도입하여 적용 및 운영하는 방안을 제시했다. 뿐만 아니라 국내 도입 시 발생할 수 있는 한계점으로 국내에서 개발한 RDB에의 적용불가, 행정정보시스템이 생산한 물리적인 전자파일의 추출 문제를 지적하고 있다.

한희정 외(2020)는 행정정보 데이터세트에 대해 법령에는 기록관리대상으로 명시하고 있으나, 구체적인 방법은 없는 것을 지적하며 데이터세트의 특성을 고려한 보존포맷이 필요하다고 한다. 이 연구에서는 데이터세트 관점에서 5가지 필수보존속성(SP, Significant Properties)인 Appearance, Behavior, Content, Context, Structure를 설정하고 이를 기반으로 다양성, 관계성, 복잡성, 이질성, 상호작용성의 5가지 데이터세트의 특성을 도출하였다. 또한 데이터세트의 5가지 특성에서 일반화, 수용성, 활용성의 3가지 고유기준을 설정하였다. 고유기준을 바탕으로 SIARD라는 보존포맷을 검증하고 재난안전 관련 공공기관의 행정정보시스템에서 수집한 RDB형 데이터세트를 SIARD로 변환하고 복원하는 검증시험을 진행하였다. 최종적

으로 SIARD는 보존포맷으로서 활용은 가능하나 지속적인 검증이 필요하다는 결과를 도출하였다.

윤성호, 이정은, 양동민(2021)은 행정정보 데이터세트 보존포맷으로 제안된 SIARD가 보존포맷으로서 정상적으로 기능할 수 있는지에 대한 기능 및 기초시험을 수행하였다. 특히 한희정 외(2020)가 제안한 데이터세트 보존포맷을 위한 3가지 고유기준인 일반화, 수용성, 활용성의 부합여부를 SIARD가 충족하는지 확인하였다. 연구 결과로 각 DBMS에서 타 DBMS 또는 SIARD 포맷으로 변환하는 과정에서 DBMS와 SIARD간 대응되는 데이터타입이 저장하는 데이터범위가 일치하지 않음으로써 일부 데이터가 누락되는 경우를 확인하였다.

국가기록원(2019)에서는 2019년 국가기록관리・활용기술 연구개발 사업으로 데이터세트 유형 전자기록의 장기보존기술 연구를 진행하면서 SIARD 2.1 표준과 SIARD Suite의 보존포맷으로서의 적절성 및 수정・보완사항을 도출하여 SIARD 기반으로 데이터세트 보존 방안을 제안하였다. 연구에서는 SIARD의 단점으로 SQL:2008 표준을 벗어나는 사항을 포함하지 못한다는 점, 자바 프로그램 안에서 SQL을 실행하기 위해 데이터베이스를 연결해주는 응용프로그램 인터페이스인 JDBC에 의존하고 있기 때문에 일부 DBMS의 고유한 요소 및 기능이 소실될 수 있다는 점, 이로 인하여 DBMS의 보존과 복원에 한계가 있다는 점을 지적하고 있다.

이어서 국가기록원(2020b)에서는 2020년 행정정보 데이터세트 기록관리 체계구축사업에서 행정정보 데이터세트 기록관리 지원도구를 고도화하였다. 사업결과물로 SIARD Suite를 기반으로 이관대상 행정정보 데이터세트를 XML 구조의 패키지 파일 형태로 추출하는 한국형 버전의 SIARD_KR이 개발되었다. SIARD_KR에서 추가된 주요 기능으로는 국산 DBMS인 큐브리드를 지원하고, BLOB 이외의 첨부파일 추출이 가능해졌으며, DB뿐만 아니라 테이블 단위 추출이 가능해졌다.

선행연구에 따르면, 행정정보 데이터세트는 현행 전자기록과는 다른 접근방식을 취해야 할 정도로 기존의 기록과는 특성이 다르며 이로 인하여 연구자들 간의 의견이 갈리고 있다. 조은희, 임진희(2009)의 경우 데이터세트를 식별하는 데 있어서 데이터베이스의 구조 혹은 그 자체를 기준으로 하는 것은 업무를 표상하기에 부적절하다고 한다. 그렇기 때문에 데이터베이스의 물리적 구조를 분석하거나 시스템 혹은 업무를 분석하여 데이터세트를 식별하려는 시도를 하였다. 반면 왕호성, 설문원(2017)은 기록의 식별 단위가 다양해지면 기록관리 행위를 설계하기 어렵고, 질서정연한 상태의 데이터베이스를 기록선별 단위에서 분리하는 것은 기록의 무결성을 해칠 우려가 있음을 밝혔다. 또한 국가기록원(2020b)의 연구결과에서는 행정정보 데이터세트의 최소단위를 데이터베이스의 구조와 일치시켜 SIARD_KR을 개발하였다.

또한 SIARD 포맷과 SIARD Suite 및 SIARD_KR의 경우 RDB의 보존포맷으로서 활용될 수 있는지에 방점이 찍혀 있었다. 다시 말해 얼마만큼 RDB가 가지고 있던 데이터를 누락 없이 보존하는가에 대한 검증이 주를 이루었고 행정정보 데이터세트를 의미 있는 정보로서 보

존하기 위한 보존포맷으로서의 적합성에 대한 내용이 부족했다.

### 2.3 시사점

데이터세트는 데이터베이스 전체를 의미하지 않는다. 데이터의 용도에 따라 하나의 데이터베이스에서도 여러 개의 데이터세트가 식별될 수도 있다. 왜냐하면 데이터베이스는 논리적 구조와 실제 물리적으로 저장되는 구조가 항상 일치하지 않기 때문이다. 오세라, 이해영(2019)의 논문에서는 하나의 업무에서 발생하는 정보가 하나의 데이터베이스에 저장될 수도 있지만, 여러 테이블에 분산되어서 저장되고 있음을 확인했다.

황진현, 백영미, 임진희(2021)의 연구에서는 KR전자조달 시스템의 데이터세트 기록을 식별하면서 테이블의 하위계층인 컬럼단위로 5개의 고유 데이터세트와 3개의 공통 데이터세트를 식별하였다. 즉 하나의 테이블 안에도 여러 데이터세트가, 혹은 여러 데이터세트의 일부가 존재하고 있음을 알 수 있다.

하지만, SIARD에서는 반드시 데이터베이스 단위로만 추출이 가능하기 때문에, 데이터의 일부분만 필요한 경우 데이터를 RDB형태로 전체를 복원한 뒤에 SQL문을 사용하여 원하는 데이터를 추출하여야 한다. 이는 국가기록원(2019)도 SIARD Suite가 모든 테이블을 대상으로 추출을 진행하고, 일부의 테이블이나 컬럼만을 추출 혹은 복원할 수 없기 때문에 전체 데이터세트를 반드시 추출해야 하는 문제점을 지적하고 별도의 기능이 추가되어야 함을 지적한 바 있다.

종합했을 때, 데이터세트라는 기록의 유형은 하나의 데이터베이스, 테이블, 컬럼이 될 수도 있지만, 시스템 목적상 여러 데이터베이스, 테이블, 컬럼에 걸쳐서 하나의 데이터세트가 식별될 수도 있다는 것을 고려하여야 한다.

또한 SIARD의 구조는 기본적으로 하나의 RDB를 해당시점에서 스냅샷하여 추출하는 기능밖에 없으므로, 하나의 테이블에서 의미 있는 정보가 모두 들어있는 경우가 아니라면, 필요 없는 데이터까지 함께 추출되는 문제점이 있다. 시스템의 상황에 맞게 데이터세트를 식별하고 추출할 수 있어야 하는데 무조건 전체의 DB를 추출해야 하는 SIARD의 구조는 중복 데이터의 문제 또한 야기한다. 여러번이나 주기적으로 DB를 SIARD로 추출하는 것을 가정할 때, 과거시점의 SIARD 파일과 현재시점의 SIARD 파일에서 무엇이 어떻게 중복되는지 확인할 수 있는 방법이 없다.

또한 SIARD는 원래의 시스템이 존재한다는 가정하에 DB에 대한 데이터만을 SIARD 포맷으로 추출하는 것이므로 원래의 시스템이 소실된다면 추출된 데이터가 어떠한 방식으로 보여지거나 혹은 서식을 갖는지는 확인할 수 없다. 뿐만 아니라 특정 시스템의 경우 이용자가 보는 UI 혹은 서식이 데이터에 의미를 부여하는 경우 SIARD에서 그것을 포착할 방법이 없다. 예를 들어서 한 학생의 정보는 단순히 표로 표현되면 의미가 없지만, 그것이 졸업증명서라는 서식으로 표현되게 되면 또 다른 특별한 의미를 가지게 된다. 하지만 SIARD에서는 서식이나 UI를 함께 보존하는 기능을 제공하지 않고 있다.

또한 BLOB과 같은 이진형태로 첨부파일을 저

장하는 데이터타입이 아니면, 첨부파일을 SIARD 포맷으로 추출하는 것 또한 불가능하다. 현재 국내 DB서버에서는 첨부파일을 저장할 때 파일시스템에 저장되어 있는 파일 경로를 DB에 저장하고, 파일 자체는 파일시스템의 경로에 저장하는 방법을 주로 사용하고 있다(국가기록원, 2019). 이러한 경우 SIARD 포맷으로 추출할 수 없어 보존포맷으로서 기능하기에 취약한 부분이라고 볼 수 있다.

마지막으로 SIARD Suite는 JDBC를 이용하여 DB에 접속하여 추출 및 복원을 진행하기 때문에 정해진 JDBC가 지원하는 6종의 RDBMS가 아닌 국내에서 점유율이 높은 큐브리드와 같은 미지원 DBMS에 대한 추출이 불가능하다. 이것은 한국형 행정정보 데이터세트 이관도구인 SIARD_KR이 개발이 요구되는 직접적인 이유가 되기도 했다.

## 3. 현행 이관도구의 한계

이후의 논의는 국가기록원에서 배포한 '행정정보 데이터세트 기록관리 지원도구 설치 파일 및 매뉴얼'의 샘플 SIARD 파일 중 'sfdbaw.siard'을 이용하여 진행하고자 한다. sfdbaw.siard는 Microsoft SQL Server 11.00.6248 버전의 DBMS로 작성된 SIARD 파일이다. 해당 샘플 파일은 스위스 국립 기록원에서 배포하는 SIARD 파일이기도 하며, 해당 데이터베이스의 원본은 Microsoft SQL Server의 실습용으로 배포되고 있는 'AdventureWorksLT'이다. 해당 샘플은 GitHub에서 배포되고 있으며 그중에서도 구문분석된 경량의 데이터베이스 샘플이다.

실제 행정정보시스템에서 사용하는 DB는 아니지만, 국가기록원에서 샘플로 제공하는 SIARD 파일이라는 점에서 누구나 접근할 수 있다는 장점이 있으며, 실제 실습용으로 사용되는 구조와 목적이 명확한 샘플DB이므로 이 SIARD 파일의 내용을 바탕으로 이후의 논의를 진행하고자 한다.

샘플은 주소, 고객, 상품, 주문정보를 관리하기 위한 데이터베이스로 총 10개의 테이블로 구성되어 있으며, 평균적으로 테이블당 9개의 컬럼과 약 427건의 데이터를 가지고 있다. 다만, 행정정보시스템과 같이 특정한 목적을 가지는 시스템에서 사용 중인 데이터베이스가 아닌 실습용 샘플 데이터베이스이므로 데이터세트를 식별함에 있어서 고객의 정보와 상품정보 그리고 주문정보를 관리하는 업무를 가정하였다.

| Row | Table name | Columns | Data records |
|---|---|---|---|
| 1 | Address | 9 | 450 |
| 2 | Customer | 15 | 847 |
| 3 | CustomerAddress | 5 | 417 |
| 4 | Product | 17 | 295 |
| 5 | ProductCategory | 5 | 41 |
| 6 | ProductDescription | 4 | 762 |
| 7 | ProductModel | 5 | 128 |
| 8 | ProductModelProductDesc | 5 | 762 |
| 9 | SalesOrderDetail | 9 | 542 |
| 10 | SalesOrderHeader | 22 | 32 |

〈그림 1〉 샘플DB의 테이블, 컬럼정보

### 3.1 현행 이관도구 SIARD_KR

SIARD_KR의 개발 배경은 다음과 같다. 철-건 구조의 표준전자문서 중심으로 설계되어 있는 현재의 기록관리 체계를 확장시켜 행정정보 데이터세트를 체계 안으로 편입시키려는 필요성과, 이에 따른 최적화된 보존포맷 선정이 요구되

었다. 특히 데이터세트는 상용화된 RDBMS를 사용하는 경우 장기보존에 적합하지 않아 다양한 RDBMS와 호환가능한 공개표준의 보존포맷이 필요했다. 국가기록원에서는 SIARD Suite가 보존포맷으로서 적절한가에 대해서 4가지 DBMS(MySQL, Oracle, SQL Server, 큐브리드)를 검증하고, 이에 따른 SQL:2008 표준사용, JDBC 사용, 외부 저장파일 추출불가, 전체 DB만을 추출 및 복원하는 등의 8가지 문제점을 지적하였다(국가기록원, 2019).

SIARD_KR은 스위스 연방기록원에서 2008년 개발 이후 공개 및 지속 보완한 SIARD Suite를 기반으로 2019년에 국가기록원이 국내환경에 맞게 국산 DBMS인 큐브리드 추출 및 복원 기능과 외부 첨부파일 추출기능 및 DB 테이블 선택기능 등을 보강하는 형태로 연구개발했고, 이관 대상 행정정보 데이터세트를 XML구조의 패키지 파일 형태로 추출하는 한국형 버전의 SIARD이다(국가기록원, 2020c).

SIARD_KR이 가지는 가장 큰 특징은 오세라, 이해영(2019)이 제안한 것처럼 데이터세트 식별의 현실성을 고려하여 추출의 최소단위를 테이블로 지정하여 RDB 보존도구가 아닌 행정정보 데이터세트 보존도구로서의 변화를 시도했다는 점이다. 왕호성, 설문원(2017)은 정보시스템 내에 존재하는 모든 데이터를 관리할 필요는 없음을 KS X ISO 16175-3을 통해서 밝히고, 데이터세트 기록은 데이터베이스의 모든 데이터가 아니라 선별 과정을 통해 기록으로 선언된 것만을 지칭한다고 했다. 즉 보존가치가 있는 데이터세트만을 선별하여 기록으로서 실질적으로 관리할 수 있는 도구가 비로소 마련된 것이다.

또한 한국 실정에 맞게 현재 G클라우드 표준으로 등록되어 있는 국내 유일의 DBMS인 큐브리드를 SIARD 포맷으로 추출할 수 있도록 하기 위한 추가 JDBC와 데이터타입간 매핑 등의 추가기능을 개발하여 적용하였다. 뿐만 아니라 실질적으로 첨부파일을 파일시스템의 경로 형태로 저장하는 경우 첨부파일이 누락되는 SIARD Suite의 단점을 보완하여 해당 파일시스템에서 직접 가져와서 SIARD 포맷으로 패키징하는 기능을 보완하여 보존포맷으로서의 기능 또한 강화하였다.

그럼에도 불구하고 SIARD_KR은 RDB 보존도구인 SIARD Suite의 장단점 대부분을 공유한다. SIARD 표준을 만족하는 완전히 새로운 프로그램을 개발한 것이 아닌, 기존에 배포된 SIARD Suite에 추가 기능을 위한 모듈을 추가한 형태로 개발되었기 때문이다. 따라서 SIARD Suite에 비하여 일부 보완되었지만 여전히 행정정보 데이터세트를 위한 보존도구로서는 부족하다고 생각된다. 이러한 SIARD_KR의 한계를 선행연구에서 이미 밝혀온 행정정보 데이터세트의 특성으로 인하여 기록으로서 관리하기 위해서 고려되어야 하는 요건들을 통해서 지적하고자 한다.

### 3.2 데이터세트 식별 단위

SIARD_KR의 경우 모체가 된 SIARD Suite보다 추출조건이 정교화되었다. RDB를 통째로 스냅샷하는 방식뿐만 아니라 원하는 테이블을 선택하여 추출할 수 있는 기능이 추가되었기 때문이다. 하지만 추출의 최소단위가 테이블인 것은 데이터세트의 특성을 모두 반영하는데 부

족하다. 그 이유는 행정정보 데이터세트를 관리하기 위해서 설정한 단위기능이라는 개념에 기인한다.

단위기능은 이규철(2016)의 연구에서 처음 언급된 개념으로 전통적인 철, 건 구조를 적용할 수 없는 행정정보 데이터세트의 특성을 고려하여 데이터세트의 이관 및 보존, 관리 단위로 제시되었다. 균일한 데이터세트를 대상으로 하는 최소 기능 단위 중 읽기를 제외한 생성, 수정, 삭제, 연산이 일어나는 기능을 의미한다. 이때 단위기능은 시스템 전체일 수도 있고, 일부 업무기능에 연결된 테이블 단위 또는 테이블의 행과 열에 저장된 데이터요소일 수도 있다. 또한 오세라, 이해영(2019)은 단위기능을 도출하는 방법으로 기능 분석 및 데이터 모델 분석을 제시하고, 데이터세트의 관리 단위는 시스템과 데이터세트의 특성에 따라 유연하게 설정되어야 함을 주장한 바 있다(서지인, 2020).

이러한 단위기능이라는 개념과 데이터세트의 특성을 고려할 때, 데이터세트는 반드시 획일적인 단위로 정의할 수 없다. 물론 KR에서 실시한 3건의 선행연구를 고려하였을 때 시스템마다 동일한 목적을 가지는 공통의 데이터세트가 존재하였다. 하지만 공통 데이터세트보다 고유한 기능을 가지는 것으로 식별된 데이터세트가 더욱 많았으며 그 형태와 크기 또한 다양했다.

특히 특정한 데이터세트의 경우 원천시스템과 원천시스템에서 데이터를 받아와서 사용하는 시스템이 존재하였다. KR의 사례에서는 통합 데이터베이스를 사용하여 밀접한 관계를 가지는 두 시스템(CPMS, EPMS)을 통합하여 데이터세트를 식별하는 방식을 사용하였다(이경남, 최광훈, 임진희, 2021). KR의 사례를 고려할 때 데이터세트로 식별될 수 있는 단위는 다양하며, SIARD_KR의 최소 추출단위가 테이블인 것은 데이터세트의 특성을 모두 반영하기가 어렵다.

고객의 배송주소를 관리하는 기능을 예로 들어보자면 〈그림 2〉~〈그림 5〉는 각각 Address, Customer, CustomerAddress, Product 테이블의 샘플데이터이다. 이 시스템은 Address 테이블에 일반적인 주소정보를 가지고 있고, Customer 테이블에 고객의 정보를 가지고 있으며, 한 고객이 여러 주소정보를 가질 수 있기 때문에 CustomerAddress 테이블을 설정하여 AddressType 컬럼에서 Main Office(메인 주소정보)와 Shipping(서브 주소정보)으로 구분하고 있다. 고객의 주소정보를 관리하려는 업무는 3개의 테이블이 모두 필요하지만, 모든 컬럼이 필요하지는 않다. 예를 들어서 모든 테이블에 공통적으로 존재하는 rowguid는 SQL Server에서 유일자 식별을 위한 고유값을 DBMS에서 자체적으로 부여하는 컬럼이다. 이는 무결성을 보장할 수 있는 기능을 할 수는 있지만, 사용자에게 보여지거나 실제로 업무에 쓰이는 데이터는 아니다. 또한 Customer 테이블에서 PasswordSalt와 PasswordHash는 각각 사용자의 비밀번호와 비밀번호의 무결성을 검증하는 해쉬값이다. 이 또한 배송정보를 관리하는데 특별한 의미를 갖지 않는 데이터이다. 따라서 3개의 테이블에서 5개의 컬럼을 제외한 나머지 컬럼들만이 하나의 '고객주소정보관리'라는 데이터세트로 식별되어야 한다.

따라서 행정정보 데이터세트의 특성을 반영

| Row | CustomerID | NameStyle | Title | FirstName | MiddleName | LastName | Suffix | CompanyName | SalesPerson | EmailAddress | Phone | PasswordHash | PasswordSalt | rowguid | ModifiedDate |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | false | Mr. | Orlando | N. | Gee | {null} | A Bike Store | adventure-\ | orlando0@ac | 245- 555- | L/Rlwxzp4w7I | 1KjXYs4= | 3F5AE95 | 01. 8. 1 오전 7:0( |
| 51 | 78 | false | Mr. | Stefan | {null} | Delmarco | {null} | Preferred Bik | adventure-\ | stefan0@adv | 819- 555- | 9rOycZRxDntX | BsI2IB4= | 8C8083/ | 01. 8. 1 오전 7:0( |
| 101 | 163 | false | Mr. | Andrew | {null} | Cencini | {null} | Sports Merch | adventure-\ | andrew2@ad | 644- 555- | Eo1pvctUvdkR | LCi+QgQ= | 989F9F3 | 03. 9. 1 오전 7:0( |
| 151 | 238 | false | Ms. | Jodan | M. | Jacobson | {null} | A Great Bicyc | adventure-\ | jodan0@adve | 652- 555- | PSKz4q56lqL3 | F/sb/Xc= | 3ECD14( | 01. 9. 1 오전 7:0( |
| 201 | 313 | false | Mr. | Dylan | {null} | Miller | {null} | Metropolitan | adventure-\ | dylan1@adve | 140- 555- | kE/AZF2xtlVSt | q1IJvKo= | 2157D9/ | 01. 11. 1 오전 8:( |
| 251 | 390 | false | Mr. | Mike | M. | Taylor | {null} | Plastic Parts ( | adventure-\ | mike6@adve | 204- 555- | 3/sdWTV1Scx/ | abgBRtE= | D6AB06: | 01. 11. 1 오전 8:( |
| 301 | 470 | false | Ms. | Delia | B. | Toone | {null} | Wingtip Toys | adventure-\ | delia0@adve | 328- 555- | UnSgG1XBHKn | k8iQbOY= | D7B637( | 03. 9. 1 오전 7:0( |
| 351 | 551 | false | Mr. | Samuel | N. | Agcaoili | {null} | Vinyl and Plas | adventure-\ | samuel0@ad | 554- 555- | jt9vdlyi0zl03w | uFYBREA= | 85A3A9 | 01. 9. 1 오전 7:0( |
| 401 | 631 | false | Mr. | Robert | P. | Lyeba | {null} | Tandem Sales | adventure-\ | robert8@adv | 631- 555- | 5obs6LI7GwH/ | /Q1sZG0= | FB9DC6( | 01. 9. 1 오전 7:0( |

〈그림 2〉 Customer 테이블의 데이터

| Row | AddressID | AddressLine1 | AddressLine2 | City | StateProvince | CountryRegion | PostalCode | rowguid | ModifiedDate |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 9 | 8713 Yosemite C | {null} | Bothell | Washington | United States | 98011 | 268AF621- 76D7 | 02. 7. 1 오전 7:00 |
| 51 | 488 | 45259 Canada W | {null} | Burnaby | British Columbia | Canada | V5G 4S4 | 6D2D6846- 3F17 | 03. 8. 1 오전 7:00 |
| 101 | 538 | 25 Danger Street | Floor 7 | Toronto | Ontario | Canada | M4B 1V5 | 10B2404C- F0AF- | 03. 8. 1 오전 7:00 |
| 151 | 588 | 99828 Routh Str | {null} | Dallas | Texas | United States | 75201 | 180F5A76- 930D | 01. 8. 1 오전 7:00 |
| 201 | 638 | 255 Irving Street | {null} | London | England | United Kingdom | C2H 7AU | D8A77424- 94AE | 03. 8. 1 오전 7:00 |
| 251 | 810 | Redford Plaza | {null} | Redford | Michigan | United States | 48239 | 43E988F4- B82D- | 01. 7. 1 오전 7:00 |
| 301 | 879 | Valley Mall | {null} | Union Gap | Washington | United States | 98903 | 860CB49B- 5774- | 02. 7. 1 오전 7:00 |
| 351 | 1009 | 2500 N Serene Bl | 19th Floor | El Segundo | California | United States | 90245 | 7619E3BD- 780B- | 03. 8. 1 오전 7:00 |
| 401 | 1059 | The Quad @ Wes | {null} | Whittier | California | United States | 90605 | 2215F750- 1863- | 03. 7. 1 오전 7:00 |

〈그림 3〉 Address 테이블의 데이터

| Row | CustomerID | AddressID | AddressType | rowguid | ModifiedDate |
|---|---|---|---|---|---|
| 1 | 29485 | 1086 | Main Office | 16765338- DBE4- | 03. 9. 1 오전 7:0( |
| 51 | 29567 | 448 | Main Office | 4F5A655A- 1259 | 03. 8. 1 오전 7:0( |
| 101 | 29643 | 899 | Main Office | 13FFDF92- 458D- | 03. 8. 1 오전 7:0( |
| 151 | 29732 | 607 | Main Office | 5D353D79- AB12 | 02. 8. 1 오전 7:0( |
| 201 | 29800 | 839 | Main Office | C0E88505- 0953- | 02. 8. 1 오전 7:0( |
| 251 | 29874 | 647 | Main Office | 8B604C39- A03E- | 03. 7. 1 오전 7:0( |
| 301 | 29939 | 475 | Main Office | EC8852E5- 975C- | 03. 9. 1 오전 7:0( |
| 351 | 30019 | 652 | Main Office | 4D79EB56- 5F62- | 02. 9. 1 오전 7:0( |
| 401 | 30098 | 486 | Main Office | 1175D1C6- 86E0 | 02. 7. 1 오전 7:0( |

〈그림 4〉 CustomerAddress 테이블의 데이터

| Row | ProductID | Name | ProductNumber | Color | StandardCost | ListPrice | Size | Weight | ProductCategoryID | ProductModelID | SellStartDate | SellEndDate | DiscontinuedDate | ThumbNailPhoto | Thu |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 680 | HL Ro | FR- R92B- 58 | Black | 1059.3100 | 1431. | 58 | 1016.04 | 18 | 6 | 98. 6. 1 오: | {null} | {null} | 0x47494638396 | nc |
| 51 | 755 | Road- | BK- R68R- 60 | Red | 884.7083 | 1457. | 60 | 8119.2( | 6 | 28 | 01. 7. 1 오: | 02. 6. 30 오 | {null} | 0x47494638396 | ro |
| 101 | 805 | LL Hea | HS- 0296 | {null} | 15.1848 | 34.2000 | {null} | {null} | 15 | 59 | 02. 7. 1 오: | 03. 6. 30 오 | {null} | 0x47494638396 | nc |
| 151 | 855 | Men's | SB- M891- S | Multi | 37.1209 | 89.9900 | S | {null} | 22 | 12 | 02. 7. 1 오: | 03. 6. 30 오 | {null} | 0x47494638396 | nc |
| 201 | 905 | ML Mc | FR- M63S- 42 | Silver | 199.3757 | 364.090 | 42 | 1274.59 | 16 | 15 | 03. 7. 1 오: | {null} | {null} | 0x47494638396 | nc |
| 251 | 955 | Tourir | BK- T79Y- 50 | Yellov | 1481.9379 | 2384. | 50 | 1153 | 7 | 34 | 03. 7. 1 오: | {null} | {null} | 0x47494638396 | ju |

〈그림 5〉 Product 테이블의 데이터

했을 때, 추출조건에 필요한 요건은 총 세 가지이다. 첫 번째로는 테이블 내부의 컬럼단위를 지정해서 추출할 수 있어야 한다. 이것은 행정정보데이터세트의 최소 식별단위인 단위기능이 반드시 테이블과 일치하지 않기 때문이다.

두 번째로는 SQL문의 WHERE절과 같이 제한조건의 역할을 하는 조건을 부여할 수 있어야 한다. 이관을 주기적으로 시행하는 경우 이미 이관된 시점 이전의 기록은 이관될 필요가 없다. 이는 불필요한 중복파일을 발생시킬 수밖에 없다. 이를 가장 쉽게 해결할 수 있는 방법으로 이관주기에 해당하는 기간을 WHERE절로 제한하면 해당하는 기간만큼의 데이터만을 추출할 수 있다.

또한 WHERE절은 제한조건으로서 기능하기 때문에 원하는 조건을 충족하는 데이터세트만을 SIARD 파일로 추출 및 보존할 수 있다는 장점이 있다. 예를 들면 〈그림 3〉의 SellEndDate 컬럼은 판매가 종료된 날짜를 의미한다. 따라서 해당 컬럼에 날짜값이 입력된 경우 해당 상품은 판매가 종료되었음을 알 수 있다. 이 경우에 "WHERE SellEndDate != "null"과 같은 조건식을 사용하여 데이터베이스 조회를 하면 판매 종료된 데이터만을 가져올 수 있다. 즉 트리거 기반의 기록을 쉽게 관리할 수 있는 기반이 된다.

마지막으로 데이터세트를 사용자가 정의하여 추출할 수 있어야 한다. 데이터세트는 물리적으로 여러 데이터베이스, 여러 테이블에 걸쳐서 존재할 수 있다. 하지만 현재의 SIARD_KR의 구조로는 한 번 추출을 실행할 때마다 하나의 SIARD 파일이 생성되게 되며, 그 과정에서 추출된 데이터베이스 혹은 테이블들이 어떠한 데이터세트로 구성되어 있는지는 알 수 없다. 따라서 논리적으로 데이터세트를 구분하여 재정의할 수 있는 옵션이 필요하다.

이미 이러한 논리적인 구분에 대한 필요성은 국가기록원(2019)에서도 하나의 물리적인 데이터베이스 전체를 보존포맷으로 변환하는 것보다 여러 개의 논리적인 데이터세트로 구분해야 한다고 주장한 바 있다. 하지만 SIARD_KR의 구조상 여러 개의 데이터세트를 각각 추출하는 경우 추출 횟수만큼의 파일이 생성되게 되며, 이는 데이터의 파편화를 야기한다. 앞서 언급한 '고객주소정보관리'라는 데이터세트를 추출하면 하나의 SIARD 파일이 생성되는데, 이 데이터세트는 데이터베이스가 가진 여러 데이터세트 중 한 가지이며, 시스템의 업무에 따라서 얼마든지 개수가 늘어날 수 있다. SIARD_KR로 데이터베이스를 데이터세트단위로 추출할 경우 원래의 DB형태로 복원할 때까지 원래의 정보가 어떠한 형태로 구조화되어 있는지 알기 어렵다. 또한 DB의 형태로 복원하더라도 데이터세트에 대한 논리적 구분에 대한 정보가 없으므로 하나의 SIARD 파일 안에 몇 개의 데이터세트가 들어있는지는 알 수 없다.

이 부분은 기록관리기준표가 보완할 수 있지만, 결국 사용자는 원래의 데이터세트에 접근하려면 RDB를 복원한 뒤에 기록관리기준표에 쓰여진 SQL문을 통해서 해당 데이터세트를 추가적으로 추출해야 한다. 따라서 SIARD_KR은 추출할 때 사전에 데이터세트를 논리적으로 나눌 수 있는 옵션이나 여러 SIARD 파일을 하나로 합칠 수 있는 기능이 필요하다.

### 3.3 중복 데이터 발생

IT 기술이 발전하면서 하나의 저장매체에 담을 수 있는 크기는 기하급수적으로 늘어났지만, 이와 함께 생산되는 데이터 또한 늘어났다. 하지만 생산되는 데이터가 늘어났다고 하는 것이 순수하게 새로운 데이터가 생산되었음만을 의미하는 것은 아니다. 예를 들어서 한 기관에서 A라는 시스템에서 생산한 기록을 B라는 시스템의 첨부파일로 첨부하는 경우를 생각할 수 있다. 이 경우 완전히 동일한 두 개의 파일을 시스템이 다르다는 이유로 모두 보존하는 것은 합리적이지 못하다.

이규철(2016)은 데이터세트 유형이 다양하여 이관, 이관 후 수집, 수집과 같이 기록관리 방법

을 다르게 해야한다고 한다. SIARD_KR의 경우 추출 시점에서 지정한 모든 테이블의 데이터를 모두 추출하는 특성이 있지만, 중복된 데이터에 대한 검증이나 제한을 하는 기능은 없다. 따라서 SIARD_KR이 사용되는 상황을 살펴보고 각각 데이터세트의 어떤 특성이 고려되어야 하는지를 살펴보려 한다.

이관도구인 SIARD_KR이 사용되는 상황은 크게 3가지로 분류할 수 있다. 첫 번째로 기관이 폐지되거나, 행정정보시스템이 사용종료되는 경우이다. 이 경우는 시스템이 보유한 데이터세트에 추가적인 변화가 예상되지 않으므로, SIARD_KR을 이용하여 행정정보 데이터세트를 이관하는 데에 적절하다. 다만, 연계되거나 혹은 원천 시스템이 존재하는 경우 해당 시스템과의 중복여부는 확인할 수 없다. 두 번째로, 기록관 혹은 타 기관으로 이관되는 경우이다. 이 상황에서는 수집과 달리, 데이터세트에 대한 관리 권한 및 권위있는 진본의 성격이 이관된 기관으로 넘어간다. 세 번째로, 타 기관에 의해 수집되는 경우이다. 첫 번째 경우와 달리 주기적으로 여러번 수집이 발생할 수 있지만, 원천시스템은 처리과에서 보유하고 있다는 점이 중요하다. 또한 이 상황에서는 행정정보시스템이 지속적으로 사용되고 있다는 점이 변수를 생성한다. 사용되는 시스템의 경우 내부의 데이터세트들이 지속적으로 변화하게 된다. SIARD_KR의 구조는 SIARD 파일을 생성할 당시의 RDB의 데이터를 스냅샷 하는 형태이므로 데이터 변화가 발생하였을 시, 그 부분에 대한 무결성을 확인하기가 어렵다.

또한 앞서 언급한 두 번째, 세 번째 이관의 경우의 수를 고려하면, 이관은 1회만 이루어지지 않을 수 있고 이것은 필연적으로 중복 데이터의 문제를 야기한다. 이는 SIARD가 구조적으로 RDB 전체를 추출하는데 초점을 두고 있고, SIARD_KR에서는 테이블 단위로 선택하여 추출하는 기능을 추가하여 필요 없는 테이블을 추출하지 않을 수 있지만, 이관되는 상황에 따른 중복파일 문제는 해결할 수 없다. 또한 여러 테이블에 걸쳐서 하나의 데이터세트로 식별되는 경우 필요 없는 컬럼까지 같이 추출되게 된다.

뿐만 아니라 선행연구에서 원천시스템의 존재로 인하여 원천시스템에 존재하는 데이터 혹은 파일이 다른 시스템에도 동일하게 존재할 수 있음을 확인한 바 있다. 이 경우 양쪽을 모두 보존하지 않고, 원천시스템의 데이터 혹은 파일을 보존하는 것이 적절하다고 판단하였다(이경남, 최광훈, 임진희, 2021). 이러한 사실을 고려할 때 중복되는 데이터를 식별할 수 있는 추가적인 기능이 필요하다.

### 3.4 서식보존의 필요성

서식의 중요성은 기록이라면 매번 강조된다. 어떠한 정보들은 서식이 없으면 그저 데이터의 나열인 기록들이 존재한다. 특히 RDB의 경우 여러 관계가 있는 표의 형태로 추상화시킬 수 있기 때문에 표 이상의 의미 있는 정보가 되기 위해서는 서식의 보존이 필수적이다. 특히 행정정보 데이터세트를 관리하려는 이유도 궁극적으로는 활용을 위해서인데, 이미 잘 활용되고 있는 서식이나 시스템의 UI를 보존하기 위한 기능은 SIARD_KR에서 고려되지 않고 있다.

이미 해당 행정정보 데이터세트를 가장 잘 활용을 하고 있는 곳은 현재 행정정보시스템을

사용 중인 처리과이며 활용을 위한 서식이나 시스템을 모두 보유하고 있다. 하지만 이미 잘 활용되고 있는 서식과 시스템이 10년, 혹은 100년이 지난 시점에서도 변하지 않으리라는 보장은 없다. 특히 완전히 시스템과 서식이 소실된 상태에서 그저 표에 지나지 않는 데이터세트를 바탕으로 원래의 시스템이나 서식을 리버스엔지니어링 하기란 불가능에 가까울 것이다. 오세라, 이해영(2019)은 데이터베이스가 원래의 시스템에서 유리되는 경우의 위험성을 이미 지적한 바 있다. SIARD_KR도 SIARD Suite나 SIARD 포맷이 가지고 있는 문제점을 그대로 답습하고 있기 때문에, 이러한 문제로부터 자유롭지 않다.

실제로 국가기록원(2020b)의 연구에서는 SIARD 포맷이 보존포맷으로서 기능할 수 있지만, 그 역할은 PDF_A와 같은 수준이고, 추가적으로 NEO 포맷으로의 추가적인 변환이 필요하다고 보았다. 즉 OAIS 참조모형에서 제시하는 CI(Content Information)나 SIP(Submission Information Package)에 해당하고, RI(Representation Information)와 함께 AIP(Archival Information Package)로의 전환이 필요하다고 보았다.

시스템을 통째로 SIARD 포맷과 같은 보존포맷으로 변환하는 방법은 없고, 그렇다고 데이터세트를 서비스해야하는 영구기록물관리기관 혹은 기록관에서 데이터세트와 관련된 모든 시스템을 복원 및 재현, 운영하기는 현실적으로 불가능하므로, 최소한 시스템이 담당하는 서식의 역할만이라도 보존할 필요성이 있다.

뿐만 아니라 한눈에 의미를 파악하기 어려운 물리적인 데이터베이스의 테이블 및 컬럼명을 실제로 표출되는 명칭으로 바꿔주는 기능도 필요하다. 왜냐하면 사용자 혹은 이용자가 실제로 보는 데이터의 명칭과 실제 데이터베이스 내부에 저장되는 컬럼의 명칭은 서로 다르기 때문이다. 데이터베이스, 테이블, 컬럼 등의 명칭은 개발당시의 명명규칙에 의하여 붙여지고, 축약어 등을 많이 이용하기 때문에 이해하기 어렵다. 뿐만 아니라 데이터베이스 내부에는 코드로 저장되지만, 실제로 표출되는 정보는 의미 있는 정보인 경우도 있다. 따라서 데이터베이스, 테이블, 컬럼 등의 실제 명칭을 표출되는, 의미있는, 이해하기 쉬운 명칭으로 논리적으로 바꿔줄 수 있는 기능이 필요하다. 〈그림 3〉을 보면 Size 필드의 경우 일반적인 숫자 사이즈와 영문 사이즈가 혼재되어 있다. 이런 경우 151번 상품의 사이즈인 'S'는 구체적으로 어떤 값에 매핑되는지에 대한 정보는 컬럼만으로 유추하기 어렵다. 이러한 상황에서 데이터베이스의 구조와 값을 추출할 경우 데이터의 품질을 낮추는 결과를 야기한다.

이 기능이 필요한 이유는 데이터세트의 활용에 있어서 중요함에도 불구하고 행정정보 시스템이 사용종료되면 획득하기 어려운 정보이기 때문이다. SIARD_KR을 통해 생성된 SIARD 포맷은 RDB의 복원으로 초점이 맞춰져 있지, 데이터세트를 관리하기 위한 포맷은 아니기 때문이다. 예를 들자면 SIARD Suite가 데이터베이스 자체만을 추출 및 복원하는 기능만을 갖고 있으며, 데이터베이스의 일부분 혹은 논리적 구분을 다루지 않음을 통해 알 수 있다.

하지만 현행 시행령 및 공공표준에 의하면, 행정정보 데이터세트의 서식정보는 필수보존정보에 해당하지는 않으므로 SIARD_KR의 기능으로는 구현되지 않은 상태이다. 하지만 데이터세트를 관리하려는 목적이 시스템의 복원이

아닌 데이터세트를 기록으로 간주하고, 장기보존 및 서비스를 하기 위함임을 고려해야 한다. 따라서 이관도구는 기록을 단순히 추출 및 보존하는 것에서 그치지 않고 추후 활용을 고려하여 서식 보존과 관련된 기능이 선택적으로 주어지는 것이 필요하다.

## 4. 행정정보 데이터세트 이관 도구의 개선방향

3장에서 살펴본 SIARD_KR의 한계점은 행정정보 데이터세트의 특성을 반영하지 못하고 있다는 공통점을 공유했다. 그러한 원인은 SIARD_KR의 모태가 RDB 보존도구인 SIARD Suite이기 때문이라고 생각한다. 데이터베이스를 통째로 보존하기 위한 도구를 그보다 더 작은 단위인 데이터세트에 대해 충분한 고려 없이 그대로 적용하였기 때문에 한계점이 뚜렷하게 드러났다. 따라서 선행연구와 앞서 살펴본 SIARD_KR의 특성 및 행정정보 데이터세트의 특성을 충분히 고려하여 SIARD_KR 뿐만 아니라, 앞으로 등장할 행정정보 데이터세트 이관도구에 적용할 수 있는 기능요건을 제시하려고 한다.

### 4.1 유연한 데이터세트 식별단위 설정

앞서 예를 든 고객주소정보관리 데이터세트를 기존의 SIARD_KR로 추출을 하면 3개의 테이블을 모두 추출해야 했다. 이때 의미 있는 정보를 구성하는 데이터가 아닌 것들까지 함께 추출하는 것을 방지하기 위하여 컬럼 단위까지 추출할 수 있는 기능을 옵션으로 부여하여야 한다. 이를 통해서 실제로 보존가치가 있는 데이터만을 추출할 수 있으며 실제 기록관리가 이루어지는 단위기능과 추출되는 단위가 일치하도록 유연하게 설정할 수 있다.

JDBC와 SQL을 사용하여 데이터베이스에서 데이터를 불러와 객체에 저장한 뒤 그것을 XML로 작성하는 SIARD_KR의 특성을 고려할 때 구현난이도도 낮고 효과는 큰 방법이다. 이러한 기능이 추가됨으로써 부가적으로 주기적 이관에 따른 중복데이터 문제를 일부 해결할 수 있다. 다만 이러한 방식으로는 무결성을 검증하기가 힘들고, 추출시점 이전의 기록이 추출시점 후에 변경된다면 그러한 내용을 추적하기가 힘들다는 단점이 있다.

SIARD_KR은 추출시점에 지정한 모든 테이블을 그대로 SIARD 포맷으로 추출할 뿐이지, 그 안에 어떠한 데이터세트들이 담겨있는지에 대한 정보는 추출하지 않는다. 뿐만 아니라 데이터베이스의 컬럼명은 명명규칙에 따라 축약어 등을 많이 사용하여 의미가 명확히 전달되지 않는 경우가 많다. 따라서 추출할 때에 어디서부터 어디까지가 하나의 데이터세트로 식별할 것인지에 대한 정보와, 컬럼이나 테이블명을 보다 더 실제로 표출되는 의미 있는 어휘를 함께 보존할 수 있는 기능이 필요하다.

또한 테이블단위로 SIARD 포맷의 내용을 보여주는 기능 뿐만 아니라 논리적으로 구분되는 데이터세트를 기준으로 하는 값과 의미있는 컬럼명을 보여주는 추가기능 또한 필요하다. 기존의 SIARD_KR은 데이터베이스의 테이블을 조회하는 것처럼 하나의 테이블에 있는 컬럼명과 데이터를 보여주지만, 이것은 데이터베이스 관점에서 데이터를 보여주는 것이므로 실

제 활용을 위해서는 사전에 식별하여 관리하는 데이터세트를 기준으로 어떠한 데이터들이 위치하는지, 그리고 어떠한 명칭으로 데이터가 표출되는지도 확인시켜줄 필요성이 있다. 따라서 SIARD 파일을 조회할 때 논리적인 데이터세트와 물리적인 데이터베이스를 선택하여 볼 수 있는 기능이 포함되어야 한다.

### 4.2 중복데이터 제거 방향

해시함수를 통해 생성되는 해시값을 통하여 데이터세트의 무결성을 보장하려는 제안은 이미 선행연구로 김주연(2020), 충남대학교 산학협력단(2015)이 연구한 내용이 존재한다. 하지만 해시함수의 고유한 특성을 이용하면 무결성뿐만 아니라 중복 데이터 문제 또한 해결이 가능하다. 다만 다른 연구에서는 데이터세트 추출의 결과로 생기는 SIARD 파일 자체의 무결성에 집중했다면, 중복 데이터 문제를 해결하기 위해서는 하나의 행에 대한 하나의 해시값을 부여하여 중복파일에 대한 검증을 할 수 있도록 하는 것을 제안한다.

해시값이란 특정한 데이터를 고정된 길이의 데이터로 매핑하는 해시함수를 통해 생성되는 값이다. 가장 큰 특징은 동일한 해시함수를 사용할 경우 입력되는 데이터가 동일하다면 동일한 해시값을 반환한다. 이는 입력되는 데이터가 다르면 무조건 다른 해시값을 반환한다는 특성을 사용하여 무결성을 검증하는 데에 주로 이용되었다. 즉 해시값이 같다면 동일한 데이터임을 알 수 있다는 것에 착안하였다.

방법은 다음과 같다. 하나의 행의 모든 컬럼을 입력값으로 해시값을 생성하여 해시테이블에 저장한다. 두 번째로, SIARD 파일을 추출할 때 이전시점의 SIARD 파일이 존재한다면 해당 파일의 해시테이블을 가져와서 참조할 수 있도록 한다. 마지막으로 이전시점과 현재시점의 해시테이블을 비교하여 존재하지 않는 행에 대해서만 데이터세트 추출을 실행한다.

해시값을 생성하는 기준을 행으로 지정한 것은 중복파일 문제와 함께 무결성 또한 보장하기 위한 수단으로 사용하기 위함이다. 주기적으로 여러 번 이관을 시행해야 하는 경우 이전시점과 현재시점 간의 데이터의 차이가 발생할 경우 해시값을 통해서 변경이 발생하였음을 인지할 수는 있지만 어느 부분에 있어서 변경이 발생한 부분을 추적할 수는 없었다. 행을 기준으로 해시값을 생성하여 이전의 데이터와 1:1 비교를 통해서 중복값이 아닌 행만을 SIARD 파일로 추출한다면 중복데이터 문제를 해소할 뿐만 아니라 어느 행에서 변경사항이 발생하였는지 또한 추적할 수 있다.

또한 첨부파일의 경우 하나의 행정정보시스템 혹은 업무관리시스템에서 생산된 기록이 다른 시스템의 첨부파일로 저장되면서 중복파일의 문제를 야기하게 된다. 대표적으로 원천시스템이 존재하는 경우가 있다. 따라서 첨부파일은 일반적인 행의 해시값과는 별도로 따로 첨부파일을 위한 해시테이블을 관리하여 원천시스템의 첨부파일 해시테이블과 비교함으로서 중복파일을 감소시킬 수 있다.

해시값 등을 이용한 중복 데이터 혹은 파일을 식별하는 기능이 이관도구에 포함되는 것은 RDB를 사용하고 있는 행정정보시스템에 대한 추가적인 설계변경 없이 가능하다는 장점이 있다. 하지만, SIARD 파일을 생성하는 과정에서

해시값을 생성하는 과정에서 추가적인 시간소요가 발생하는 점과, 해시테이블이 이관도구에 저장됨으로 인하여 타 기관의 중복 데이터 및 파일은 식별할 수 없다는 단점은 보완되어야 할 것이다.

### 4.3 룩앤필을 위한 서식보존

기록은 내용, 구조, 맥락의 3가지 요소로 구성된다. 전자기록물 이전에는 이러한 요소가 결합된 종이문서에서는 이러한 룩앤필을 고려할 필요가 없었다. 하지만 전자기록의 경우 디스플레이되는 인터페이스로 기록의 3가지 요소를 확인할 수 있으며 따라서 재현성이 중요하다(왕호성, 설문원, 2017).

CCSDS(2012)에서 제안하는 디지털 정보의 장기보존을 위한 기술적 권고사항을 제시하는 OAIS 참조모형에서는 시스템의 접근과 활용 서비스의 보존을 위하여 액세스 소프트웨어의 룩앤필 보존이 중요하다고 한다. OAIS 참조모형에서는 포팅, 에뮬레이션 등의 방법으로 원래 사용하던 시스템을 그대로 구현하여 원래의 룩앤필을 보존하는 방법을 제안하고 있다. 특히 행정정보 데이터세트의 경우 업무는 행정정보시스템에서 이루어지고 업무의 결과가 데이터베이스에 저장되는 구조인데, 이러한 과정에서 업무가 어떻게 진행되었는지에 대한 과정은 알 길이 없다. 특히 행정정보시스템이 사용종료되거나 고도화 사업으로 인하여 이전의 시스템이 다른 시스템으로 통폐합되는 경우 이용자에게 보여지는 인터페이스가 크게 바뀔 가능성이 크다. 또한 바뀐 환경에서 SIARD_KR로 추출한 데이터를 복원했을 때 바로 원래의 상태로 돌아갈 수 있으리라는 보장 또한 할 수 없다. 이와 같이 의미 있는 정보를 구성하는 요소로 서식을 빼놓을 수 없으며 이를 함께 보존하는 것 또한 중요하다.

따라서 서식을 보존하기 위한 방법으로 가장 간단하게 서식정보를 캡처하는 것을 생각해 볼 수 있다. 데이터세트의 내용보다 서식이 더 중요하다고 식별된 데이터세트에 대하여 서식의 캡쳐를 함께 추출할 수 있는 기능이 필요하다. 따라서 추출할 데이터세트를 지정할 때 옵션으로 서식의 캡쳐를 함께 저장할 수 있는 추가 기능이 주어져야 한다.

또 다른 방안으로는 캡처 대신 웹 페이지 자체를 서식정보로 취급하여 같이 보존하는 것이다. 웹 애플리케이션의 기본적인 구조는 클라이언트로부터 요청이 들어오면 필요한 데이터를 데이터베이스로부터 가져와서 지정된 처리를 거친 후 웹페이지 위에 결과값을 보여준다. 웹 기반으로 제공되는 행정정보시스템은 이러한 HTML로 쓰여진 웹페이지를 모두 가지고 있으며, 따라서 그 웹페이지를 함께 보존하고, 처리의 결과가 표출되는 태그를 데이터세트의 컬럼명과 연결 지어주면 앞서 언급한 캡처의 방법보다 더 효과적인 룩앤필을 보존할 수 있는 수단이 된다. 따라서 캡처 외에도 .xml 및 .html확장자에 대한 보존이 옵션으로 주어져야 한다.

앞선 캡처의 방법보다 웹페이지 자체를 보존한다면 실제 실무자가 사용하는 UI를 그대로 표현할 수 있다는 장점과 원래의 서식에 데이터세트값을 바로 랜더링 해주는 뷰어를 구현할 수 있다는 장점이 있지만, 웹 기반의 시스템에만 한정되고, 실제 시스템을 구현하는 것이 아니기 때문에 서식 등의 UI를 웹페이지로 구현

한 형태의 웹기반 행정정보시스템에만 적용할 수 있다는 한계점이 있다.

앞선 두 가지 개선방안은 SIARD 포맷 내부에 서식과 UI에 해당하는 부분을 보존하는 방법이라면, 보존포맷을 이용하는 것은 SIARD 포맷의 한계점을 보완하기 위하여 다른 보존포맷을 이용하는 방법이다. SIARD 포맷의 한계점은 한희정 외(2020)가 이미 밝힌 바 있으며, 국가기록원(2020b)의 연구결과에서도 SIARD를 NEO로 다시 인캡슐레이션하는 방안을 짧게 제안하였다.

앞선 두 가지 방법이 SIARD파일 내부에 서식관련 파일을 함께 인캡슐레이션 하는 방법이라면, 마지막 방안은 SIARD 포맷 이외의 다른 보존포맷을 이용하여 서식정보를 인캡슐레이션하는 방법이다. 따라서 SIARD_KR의 기능개선이나 SIARD 포맷의 기술요소를 변경하지 않아도 서식정보를 보존할 수 있는 방법이며, 위에서 제안한 캡처나 혹은 웹페이지 자체를 기록관리 메타데이터로 포착하여 SIARD 파일과 함께 인캡슐레이션하여 시스템이 존재하지 않는 상태에서도 일부 룩앤필을 보존할 수 있는 방안이다. 또한 데이터베이스의 크기가 너무 큰 경우와 같이 부득이하게 하나의 데이터베이스를 여러 SIARD 파일로 추출할 경우, 파편화되지 않도록 할 수도 있다.

## 5. 결 론

SIARD_KR은 스위스 국립 기록원이 개발한 RDB 보존도구인 SIARD 포맷 및 SIARD Suite에 기반하여 개발되었지만 선행연구에서 행정정보 데이터세트의 특성으로 제시된 사항에 대해서 충분한 고려가 이루어지지 않은 채 국내로 도입되었고 실제로 기능을 살펴보니 행정정보 데이터세트 이관도구가 아닌 한국형 DBMS인 큐브리드를 지원하는 RDB 보존도구에 그치는 기능을 보여주었다. 행정정보 데이터세트의 특성이 반영되지 않은 이관도구는 중복 데이터 및 파일의 발생, 행정정보시스템의 룩앤필 및 서식의 보존불능 등의 문제를 야기할 것으로 예상된다.

진본성, 신뢰성, 무결성, 이용가능성을 가지는 정보를 기록이라고 부르는 것처럼 공공기관이 보유한 모든 데이터를 기록이라고 부를 수는 없다. 행정정보 데이터세트도 기록으로 선언된 이상 이러한 대명제에서 자유로울 수 없다. 그런데 보존을 위해 이관하는 과정에서 행정정보 데이터세트의 특성이 일부 소실된 채 단순한 RDB의 형태로서 이관되어서는 안된다고 생각한다.

특히 공공데이터에 대한 관심이 집중되는 상황에서 행정정보 데이터세트를 기록이 아닌 RDB의 형태로 이관받겠다는 것은 이해할 수 없는 결정이다. 단순히 DB안에 들어있는 데이터를 다루고 분석하는 것은 데이터 전문가의 영역이다. 기록이 기록다워질 수 있는 것은 행위를 표상하는 것에 있다고 생각한다. 그리고 기록이 행위를 표상할 수 있도록 구조화하는 것은 기록전문가의 영역이라고 생각한다. 행정정보 데이터세트가 그냥 데이터의 집합이 아니라 진정으로 의미 있는 정보가 되기 위해서는 행위를 표상할 수 있을 만큼의 충분한 재현정보가 원래의 데이터와 함께 하여야 한다고 생각하고 그에 따른 SIARD_KR의 개선방안을 정리하고 추후 다른

이관도구에도 적용할 수 있도록 기능요건처럼 제시해 보았다.

지금도 체계적인 관리를 받지 못하는 행정정보 데이터세트들이 관리의 어려움을 이유로 은근슬쩍 폐기되거나 이용가능성은 배제된 채 그저 저장매체에 쌓여가고 있을지도 모르는 상황에서, 이 논문이 행정정보 데이터세트가 실질적으로 관리되는데 조금이나마 기여할 수 있기를 기대하는 바이다.

## 참 고 문 헌


국가기록원 (2019). 데이터세트 유형 전자기록의 장기보존기술 연구.

국가기록원 (2020. 12. 07). 행정정보 데이터세트 기록관리 지원도구 설치 파일 및 매뉴얼. 샘플 SIARD 파일(sfdbaw.siard). 출처: https://www.archives.go.kr/next/manager/infoDataDetail.do?board_seq=97620&page=1&keytype=&keyword=

국가기록원 (2020a). NAK 35 : 2020(v1.0) 행정정보 데이터세트 기록관리 기준 - 관리기준표 작성 및 이관규격(v1.0).

국가기록원 (2020b). 2020년 행정정보 데이터세트 기록관리 체계구축 사업 결과보고서.

김주연 (2020). SIARD를 활용한 행정정보 데이터세트 장기 보존방안 연구. 석사학위논문, 명지대학교 기록정보과학전문대학원 기록관리학과.

노명환 (2020). 4차 산업혁명 시대 데이터 아카이브와 기록관리의 길: 국가기록원의 행정정보 데이터세트 기록관리 실행방안에 대한 비판적 검토와 중·장기 차원의 제안. 기록과 정보·문화 연구, 1, 7-43. https://data.doi.or.kr/10.23035/kaics.2020.1.11.007

류한조, 백영미, 임진희 (2021). 데이터세트 생산시스템 기능요건 연구: KR 재산관리시스템 사례를 중심으로. 기록학연구, 70, 5-40. https://doi.org/10.20923/KJAS.2021.70.005

서지인 (2020). 행정정보 데이터세트 관리 개선방안 연구: 공공데이터와의 비교를 중심으로. 한국기록관리학회지, 20(4), 41-58. https://doi.org/10.14404/JKSARM.2020.20.4.041

송치호, 임진희 (2022). 행정정보데이터세트의 데이터 품질평가 연구. 기록학연구, 71, 237-272. https://data.doi.or.kr/10.20923/kjas.2022.71.237

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

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

왕호성, 설문원 (2017). 행정정보 데이터세트 기록의 관리방안. 한국기록관리학회지, 17(3), 23-47.

https://doi.org/10.14404/JKSARM.2017.17.3.023
윤성호, 이정은, 양동민 (2021). 행정정보 데이터세트 보존포맷으로서 SIARD 검증에 관한 연구. 한국기록관리학회지, 21(3), 99-118. https://doi.org/10.14404/JKSARM.2021.21.3.099
이경남, 최광훈, 임진희 (2021). 데이터세트 기록관리를 위한 기록관의 역할 연구: KR 사업관리시스템 사례를 중심으로. 정보관리학회지, 38(3), 263-285. https://doi.org/10.3743/KOSIM.2021.38.3.263
이규철 (2016). 행정정보시스템 데이터세트의 이해와 기록관리 고려사항. 기록관리 표준・거버넌스 포럼 자료집, 72-78.
임진희 (2021). 공문서의 기계가독형(Machine Readable) 전환 방법 제언. 기록학연구, 67, 99-138. https://doi.org/10.20923/KJAS.2021.67.099
조은희, 임진희 (2009). 행정정보 데이터세트 기록의 선별 기준 및 절차 연구. 기록학연구, 19, 251-291. https://doi.org/10.20923/KJAS.2009.19.251
충남대학교 산학협력단 (2015). 데이터세트 구조분석 및 진본성 보장 기록관리 기능모델 연구. 대전: 국가기록원. https://doi.org/10.23000/TRKO201600002968
한희정, 윤성호, 오효정, 양동민 (2020). 데이터세트 보존포맷 검증방안에 관한 연구: 재난안전정보 데이터세트의 SIARD 적용을 통해. 정보관리학회지, 37(2), 251-284. https://doi.org/10.3743/KOSIM.2020.37.2.251
황진현, 백영미, 임진희 (2021). 공공기관 데이터세트 식별과 평가 절차 연구: 국가철도공단 전자조달시스템 사례를 중심으로. 기록학연구, 70, 41-83. https://doi.org/10.20923/KJAS.2021.70.041
Consultative Committee for Space Data Systems (2012). Reference Model for an Open Archival Information System (OAIS).

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

Cho, Eun-Hee & Yim, Jin-Hee (2009). A study on record selection strategy and procedure in dataset for administrative information. The Korean Journal of Archival Studies, 19, 251-291. https://doi.org/10.20923/KJAS.2009.19.251
Chungnam National University (2015). A Research on Dataset Structure Analysis and Record Management Functional Model for Assurance Record Authority. Ministry of the Interior and Safety. https://doi.org/10.23000/TRKO201600002968
Han, Hui-Jeong, Yoon, Sung-Ho, Oh, Hyo-Jung, & Yang, Dongmin (2020). Empirical verification of conversion and restoration of preservation format for dataset: application of dataset with

disaster safety information to SIARD. Journal of the Korean Society for Information Management, 37(2), 251-284. https://doi.org/10.3743/KOSIM.2020.37.2.251

Hwang, Jin-Hyun, Baek, Young-Mi, & Yim, Jin-Hee (2021). Study on public institution dataset identification and evaluation process: focusing on the case of KR electronic procurement system. The Korean Journal of Archival Studies, 70, 41-83. https://doi.org/10.20923/KJAS.2021.70.041

Kim Ju-Yeon (2020). A Study on the Long-term Preservation of Administrative Information Datasets Using SIARD. Master thesis, Graduate School of Records, Archives & Information Science, Myongji University.

Lee, Kyuchul (2016). Understanding of the administrative information system dataset and considerations for record management. Record Management Standard · Governance Forum Data Collection, 72-78.

Lee, Kyung-Nam, Choi, Kwang-Hoon, & Yim, Jin-Hee (2021). A study on the role of records center for dataset records management: focused on case study of KR project management system. Journal of the Korean Society for Information Management, 38(3), 263-285. https://doi.org/10.3743/KOSIM.2021.38.3.263

National Archives of Korea (2019). Study on Long-term Preservation Technology of Dataset-type Electronic Records.

National Archives of Korea (2020a). Record Keeping Criteria for Dataset - Composition of Dataset Management Reference Table & Exchange of Dataset - Version 1.0

National Archives of Korea (2020b). Report on the Results of the 2020 Administrative Information Data Set Record Management Project.

Noh, Meung-Hoan (2020). The way for data archive and records/archive management in the 4th industrial revolution era: critical reviews and mid- and long-term proposals for the national archives' administrative information dataset records/archive management implementation plan. The Korean Journal of Archival, Information and Cultural Studies, 1, 7-44. https://data.doi.or.kr/10.23035/kaics.2020.1.11.007

Oh, Seh-La & Rieh, Hae-Young (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, Seh-La, Park, Seung-Hoon, & Yim, Jin-Hee (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

Ryu, Han-Jo, Baek, Young-Mi, & Yim, Jin-Hee (2021). A study on the functional requirements of record production system for dataset: focused on case study of KR asset management system. The Korean Journal of Archival Studies, 70, 5-40. https://doi.org/10.20923/KJAS.2021.70.005

Seo, Jiin (2020). A study on the improvement of data set management in government information systems: a comparison with public data. Journal of Korean Society of Archives and Records Management, 20(4), 41-58. https://doi.org/10.14404/JKSARM.2020.20.4.041

Song, Chiho & Yim, Jin-Hee (2022). A study on data quality evaluation of administrative information dataset. The Korean Journal of Archival Studies, 71, 237-272. https://data.doi.or.kr/10.20923/kjas.2022.71.237

Wang, Ho-Sung & Seol, Moon-Won (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

Yim, Jin-Hee (2021). Suggestions on how to convert official documents to machine readable. The Korean Journal of Archival Studies, 67, 99-138. https://doi.org/10.20923/KJAS.2021.67.099

Yoon, Sung-Ho, Lee, Jung-Eun, & Yang, Dongmin (2021). A study on SIARD verification as a preservation format for data set records. Journal of Korean Society of Archives and Records Management, 21(3), 99-118. https://doi.org/10.14404/JKSARM.2021.21.3.099