논문

행정정보데이터세트의 데이터 품질평가 연구*

A Study on Data Quality Evaluation of Administrative Information Dataset

송치호(Song, Chiho)** · 임진희(Yim, Jinhee)***

* 본 연구는 2021년 국가기록원 “행정정보데이터세트 기록정보 서비스 및 활용모형 연구”의 결과를 바탕으로 작성되었음.
** (사)한국국가기록연구원 수석연구원(제1저자)(chihosong@gmail.com).
*** 명지대학교 기록정보과학전문대학원 조교수(교신저자)(yimjhkr@mju.ac.kr).
■투고일: 2021년 12월 31일 ■최초심사일: 2022년 01월 04일 ■최종확정일: 2022년 01월 20일.
■기록학연구 71, 237-272, 2021, https://doi.org/10.20923/kjas.2022.71.237

〈초록〉

2019년부터 국가기록원의 주도로 행정정보데이터세트 기록관리체계 구축 시범사업이 본격적으로 시작되었다. 2021년까지 3년에 걸친 사업의 결과를 바탕으로 개선된 행정정보데이터세트 관리방안이 공공기록물 관련 법령과 지침에 반영될 예정이다. 이를 통해 행정정보데이터세트는 본격적인 공공기록관리의 대상이 된다. 공공기록이 전자문서 중심으로 전환되었고 행정정보시스템의 데이터세트까지 본격적인 공공기록관리의 대상으로 포함되었지만, 기록을 구성하는 원 자료(raw data)로서의 데이터 자체의 품질 요건에 관한 연구는 아직 부족한 상황이다.

데이터 품질이 보장되지 않으면 데이터의 구성체이며 기록의 집합체인 데이터세트는 기록의 4대 속성 전체가 위협받게 된다. 더욱이 표준기록관리시스템의 규격을 고려하지 않고 기관 실무 부서의 다양한 요구를 반영하여 구축된 행정정보시스템의 데이터는 기록관리 관점에서 그 품질에 대한 신뢰성이 부족할 경우 공공기록 자체의 신뢰성을 확보할 수 없을 것이다.

본 연구는 2021년 국가기록원에서 진행한 “행정정보데이터세트 기록정보 서비스 및 활용모형 연구”에서 제시된 행정정보데이터세트 관리방안을 기반으로, 적극적으로 개념이 확장된 평가, 그중에서 데이터 품질평가에 관한 연구를 수행하였다. 범정부적으로 추진되고 있는 다양한 데이터 , 특히 공공 데이터 관련 정책과 가이드를 참고하여 기록관리 차원에서의 품질평가 요건을 도출하고, 구체적인 지표를 제시해 보고자 한다. 이를 통해 향후 본격화될 행정정보데이터세트 기록관리에 도움이 되기를 기대한다.

주제어 : 행정정보데이터세트, 데이터세트, 데이터, 공공데이터, 데이터 품질관리, 평가, 데이터 품질 평가

〈Abstract〉

In 2019, the pilot project to establish a record management system for administrative information datasets started in earnest under the

leadership of the National Archives. Based on the results of the three-year project by 2021, the improved administrative information dataset management plan will be reflected in public records-related laws and guidelines. Through this, the administrative information dataset becomes the target of full-scale public record management. Although public records have been converted to electronic documents and even the datasets of administrative information systems have been included in full-scale public records management, research on the quality requirements of data itself as raw data constituting records is still lacking.

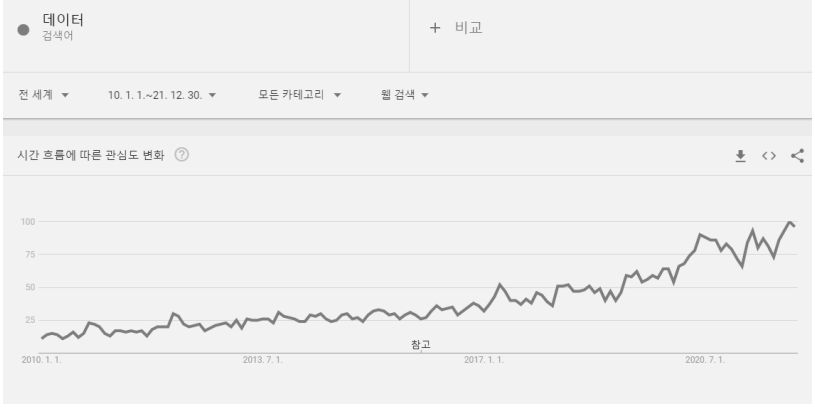
If data quality is not guaranteed, all four properties of records will be threatened in the dataset, which is a structure of data and an aggregate of records. Moreover, if the reliability of the quality of the data of the administrative information system built by reflecting the various needs of the working departments of the institution without considering the standards of the standard records management system is insufficient, the reliability of the public records itself can not be secured.

This study is based on the administrative information dataset management plan presented in the "Administrative Information Dataset Recorded Information Service and Utilization Model Study" conducted by the National Archives of Korea in 2021. A study was conducted. By referring to various data, especially public data-related policies and guides, which are being promoted across the government, we would like to derive quality evaluation requirements in terms of records management and present specific indicators. Through this, it is expected that it will be helpful for record management of administrative information dataset which will be in full swing in the future.

Keywords : administrative information datasets, dataset, data, public data, data quality management, appraisal, data quality evaluation

1. 서론

〈그림 1〉 구글 트렌드 '데이터' 탐색결과



〈그림 1〉은 2010년부터 2021년 현재까지 구글 트렌드를 통하여 '데이터'에 대한 전 세계의 관심도의 증가 추세이다. 빅데이터와 컴퓨팅능력의 도약, 그로 인한 인공지능기술의 발전 등을 통해 데이터에 관한 관심이 급격하게 증가하는 모습을 확인할 수 있다.

데이터의 관점에서 본다면 공공기록데이터에서 양적인 측면에서 절대적인 비중을 차지하고 있는 행정정보데이터세트(이하 '데이터세트')를 공공기록관리체계에 본격적으로 포함하려는 정책적 시도는 2005년 기록관리혁신 로드맵이 발표된 이후 현재까지 꾸준히 진행되었다. 이러한 과정의 일환에서 데이터세트에 대한 관리체계가 본격적으로 수립되는 것은, 2016년에서 2017년에 걸쳐 국가기록원이 데이터세트 기록관리 방안에 대한 초안을 마련하고 재설계 안까지 확정한 시점으로 볼 수 있다. 이렇게 데이터세트 기록관리 방안이 마련되고 2019년 최초로 시범사업이 시행된 이후, 데이터세트 관리체계는 데이터세트의 영역별

정보를 담고 있는 데이터세트 관리기준표를 중심으로 데이터세트의 보유 및 보존 기간을 평가하도록 기본적인 절차가 설계되었다.

데이터세트에 관련한 선행연구는 이러한 정책적인 기조에 부응하여, 전반적인 기록관리의 관점에서 데이터세트의 관리정책이나 방향을 종합적으로 제시하거나, 데이터세트 관리의 절차와 방법 혹은 데이터세트 자체의 구조와 성격을 연구하는 것으로 일별 할 수 있다.

데이터세트 관리정책과 방향에 대해서, 김유승(2018)은 "디지털시대의 공공기록평가에 관한 정책적 고찰"에서 영국 TNA 평가정책의 변화를 정책적 관점에서 논하고, 현 단계 영국 공공기록 평가체제를 분석하여 영국 공공기록 평가체제의 시사점을 바탕으로 우리나라 공공기록 평가의 정책적 개선방안을 제시하였다, 설문원(2020)은 "디지털 전환 시대의 공공기록정책"에서 영국과 호주를 사례를 분석하여 행정정보시스템의 디지털정보의 경우 보존과 활용목적을 동시에 충족시켜야 한다는 점에서 기록 자산으로서의 정보관리 접근법을 강조하며 평가정책의 변화 필요성, 데이터세트 평가 시 고려사항, 평가정책의 변화 방향을 제안하였다.

데이터세트 관리의 절차와 방법에 대해서, 조은희, 임진희(2009)는 데이터세트 기록의 기본적인 선별기준을 분석하고 선별절차를 제안하였고, 임진희, 조은희(2010)는 "행정정보 데이터세트 기록 이관 시 데이터 보정 및 품질 개선 방법 연구"에서 데이트세트 기록을 이관할 때 필요한 데이터 보정과 품질에 대한 개선방안을 연구하였다. 특히 후자의 경우 데이터웨어하우스에서 사용하는 추출방안(Extraction, Transformation, Transportation, ETT)을 방법론으로 하여 데이터 이관 시 데이터 보정절차와 품질개선 사례를 제시하였는데, 보정절차에서 제시되는 데이터세트의 볼륨 분석 유효값 확인, 코드값 및 데이터설명 정보 확보 방안 등은 현재의 데이터세트 품질평가 방법론에서도 유용하게 적용할 수 있는 방안으로 판단된다.

황진현, 박종연, 이태훈, 임진희(2014)는 “행정정보시스템 기록 이관 절차와 방법 연구”에서 기존의 원칙과 방법론에서 나아가 원자력안전위원회의 업무기록 사례를 통하여 행정정보시스템의 이관 절차를 제시하고, 기록관리 모듈의 기능과 DB설계방안을 제안하였다. 오세라, 박승훈, 임진희(2018)는 “행정정보 데이터세트 사례 조사 연구”에서 관리대상 선별, 관리 프로세스와 제도에서 행정정보데이터세트는 존재 형식, 생산 프로세스, 보존기간, 처분 방법, 관리 주체의 차원에서 현행의 전자문서와는 다르다는 것을 밝히고 각 데이터세트의 고유한 요구를 반영하여 관리해야 할 것을 제안하였다. 오세라, 이해영(2019)은 “행정정보 데이터세트의 기록관리 방안”에서 6개 공공기관 행정정보시스템 데이터세트의 현황을 조사하고, 데이터세트 관리기준의 주요 고려사항을 도출하고 데이터세트의 기록관리 기준과 절차를 제안하였다.

최근에 국가철도공단(이하 ‘KR’)의 행정정보데이터세트 사례를 중심으로 연구가 수행되었는데, 류한조, 백영미, 임진희(2021)는 “데이터세트 생산시스템 기능요건 연구: KR재산관리시스템 사례를 중심으로”에서 재산관리시스템의 기록관리기능요건을 연구하였다. 황진현, 백영미, 임진희(2021)는 “공공기관 데이터세트 식별과 평가 절차 연구: 국가철도공단 전자조달시스템 사례를 중심으로”에서 데이터 식별과 평가, 평과결과에 따른 보존 기간 책정과 이관 방안을 제시하였다.

그런데 데이터의 품질이 보장되지 않으면 데이터세트, 더 나아가 기록 자체를 신뢰할 수 없으며, 그 사용성까지 의심되지만 여기에 관한 연구는 2010년에 진행된 임진희, 조은희의 연구가 거의 유일할 정도로 데이터세트의 데이터 품질의 문제는 기록관리 연구의 대상에서 상대적으로 등한시되었고, 데이터세트 관련 시범사업에도 고려되지 않았다.

2021년에 국가기록원에서 진행한 “행정정보데이터세트 기록정보 서비스 및 활용모형 연구”(이하 ‘2021년 연구’)에서는 2년간의 시범사업 결과에 따라 제기된 기관들의 의견과 사업 경험을 반영하여 데이터세트

의 관리기준표와 관리시스템에 대한 설계안을 제시하고, 데이터세트의 평가·선별에 대한 개념 확장 및 개선방안과 이를 반영한 새로운 레코드스케쥴에 대한 연구를 수행하였다. 기존의 데이터세트 평가가 공공기록물관리의 평가 패러다임의 연장선상에서 소극적인 사후평가의 형태로 진행되었다면, 2021년의 연구는 적극적인 사전평가로 그 개념을 확장할 것을 권고하고 있고, 평가의 영역을 업무기능 및 내용 가치, 데이터 연계·결합 가치, 데이터 품질 가치로 확장하고 있다. 특히 기존의 기록관리 관점에서 데이터에 대한 관점이 데이터 기술정보로서의 메타데이터의 적합성, 장기보존을 위한 파일포맷 및 보존패키지 등에 국한되었다면 데이터 품질평가의 도입은 기록을 구성하는 구체적인 데이터 자체에 대한 품질에 대한 보장과 이를 통한 기록의 미래 재현성을 확보하는 기회가 될 것으로 예상한다.

본 연구는 새롭게 설계된 데이터세트 평가의 한 영역으로서, 데이터세트 기록의 데이터 품질에 대하여 데이터, 특히 공공데이터의 품질관리체계와의 유사점과 상이점을 분석하고, 기록관리 관점에서 데이터 품질을 평가할 수 있는 평가영역과 지표, 평가도구를 제시한다. 그리고 국가철도공단에서 진행한 데이터 품질평가의 사례를 소개하고, 앞에서 제시된 지표와 평가도구의 실제 평가과정에서의 적용내용과 고려사항을 제시하였다. 또한 평가 결과에 대한 분석과 권고를 통해 실제 시스템의 운영과 데이터품질 관리 원칙의 괴리를 확인하고 접점을 찾으려는 시도를 소개하였다. 본 연구를 통해 기존의 행정정보데이터세트 연구에서 미진했던 기록의 데이터 품질평가에 대한 출발점이 되기를 기대한다.

2. 공공데이터와 행정정보데이터세트 데이터 품질관리

1) 데이터와 데이터 품질관리의 개념

한국정보통신기술협회(이하 'TTA') 정보통신용어 사전에서는 데이터를 "인간 또는 컴퓨터를 비롯한 자동 기기에 의해 행해지는 통신과 해석, 처리로 형식화된 사실과 개념, 명령을 표현한 것"으로 정의하고 있다. 그리고 국제표준기구(이하 'ISO')에서는 ISO 8000-2:2020(en)에서 데이터를 의사소통(communication), 이해(interpretation), 절차 수행(processing)에 맞추어 정형화된 형태로서, 정보의 재해석될 수 있는(reinterpretable) 재현"으로, 데이터 품질을 "내재한 데이터 특징의 집합이 요구사항을 만족하는 정도"로 정의하고 있다.

데이터의 품질관리에 대해서는 TTA에서는 데이터 품질관리를 "데이터의 적합성, 적시성, 정확성, 무결성, 적절성 및 접근 가능성 등을 이르는 용어"로 정의하고 있으며, ISO는 "데이터 품질에 관련된 조직에 대한 지도와 통제를 수행하는 협력적 활동"으로 정의하고 있다. TTA의 정의가 전통적인 의미에 충실하다고 한다면, ISO의 정의는 데이터에 대한 최근의 범세계적 트렌드를 반영한 것으로 보인다. 종래의 DIKW(Data Information Knowledge Wisdom)피라미드에서는 데이터를 순수한 원자료(raw data)로서, 의미를 지니는 정보로 진화하기 위한 근거(basis)로 정의했지만, ISO는 데이터의 가능성에 초점을 맞추어 정보의 '재해석될 수 있는' 재현으로 정의하고 있는 것이다. 이는 인공지능에 의한 동일한 데이터에 대한 무한한 해석 가능성을 강조하는 것이며, 해석 가능성 자체가 기계학습의 과정과 결과라고 판단하고 있는 것으로 보인다. 이런 관점에서 ISO가 기존의 조직 정보기술 거버넌스(ISO/IEC 38500:2015에서 정의했던 원칙(책임, 전략, 획득, 수행 등)을 이제 데이터 영역까지 그대로 적용하도록 개념을 확대한 것이다. 데이터 품질과 데이터 품

질관리는 이제 요구사항에 대한 만족도로, 품질관리 조직에 대한 관리와 협력으로 정의되는데, 품질관리 활동은 해당 조직 내에서만 국한되는 것이 아니라, '협력적 활동'으로, 범부서적으로 추진되어야 할 활동으로 명확하게 적시하고 있다.

이렇게 본다면 전통적인 데이터 품질관리 방법론이 데이터 생산 이후에 데이터의 결측값이나 오류를 평가하는 것이었다면, 이제 해당 데이터를 소유하고 운영하는 부서와 기관이 생산 단계 이전에서부터 데이터의 요구조건과 향후 활용방안까지를 고려해서 수행되는 방향으로 그 패러다임이 전환되었다고 볼 수 있다.

2) 공공데이터 관리체계

"공공데이터의 제공 및 이용 활성화에 관한 법률"(이하 '공공데이터법')에서는 공공데이터를 "데이터베이스, 전자화된 파일 등 공공기관이 법령 등에서 정하는 목적을 위하여 생성 또는 취득하여 관리하고 있는 광(光) 또는 전자적 방식으로 처리된 자료 또는 정보"로 정의하고 있다, 그리고 대상을 전자정부법에 따른 행정정보(전자정부법 제2조6호), 지능정보화 기본법에 따른 공공기관이 생산한 정보, 공공기록물 관리에 관한 법률(이하 '공공기록물법')에 따른 전자기록물 중 대통령령으로 정하는 전자기록물로 정의하고 있다. 대통령령인 공공데이터법의 시행령에서는 전자기록물을 구체적으로 웹 기록물 및 행정정보데이터세트로 명시하고 있는데, 이는 공공기록물법과 그에 따른 기록관리체계에서의 행정정보데이터세트와 중복된다(설문원, 2019).

이러한 현상은 행정안전부가 추진하고 있는 정보자원 보존체계 구축 계획과 그 일환으로 진행된 연구(행정안전부, 2019)에 기인한 것으로서, 공공데이터가 '공공데이터의 개방'의 영역으로서 기록보존, 정보시스템 운영, 정보자원보존과 함께 행정안전부의 정보자원 보존체계에서 유관

분야 R&R 4개 영역(〈표 1〉) 중 하나를 구성하고 있기 때문이다. 정보자원 보존체계는 물리적인 실체로서 동일한 디지털 데이터에 대해서 공공데이터, 기록정보, 정보자원 등으로 개념 및 영역을 각각 정의하고 접근방법과 기반체계를 일괄적으로 통합하는 것을 목표로 하고 있으며, 이에 따르면 기록관리의 영역은 정보자원 보존체계의 하위영역으로서 '기록보존'의 영역에 국한된다.

〈표 1〉 정보자원 보존체계 유관분야 R&R

	공공데이터 개방	기록보존	정보시스템운영	정보자원보존
관련기관	행정안전부 공공데이터정책과	국가기록원 기록관	국가정보 자원관리원	행정안전부 전자정부
관련 법·제도	공공데이터의 제공 및 이용 활성화에 관한 법률/시행령/시행규칙	공공기록물 관리에 관한 법령/시행령/시행규칙 대통령기록물 관리에 관한 법률/시행령	행정안전부와 그 소속 기관 직제	전자정부 시행령
속성/특성	국가차원의 데이터 중요성 데이터 활용수준 데이터의 잠재가치	신본성 신뢰성 무결성 이용가능성 KS X ISO 15489	입주기관 정보시스템 안정적 운영 재해복구시스템 구축 데이터 백업·소산	지속성 선별성 무결성 활용성
관련 정보시스템	공공데이터 개방포털	기록관리시스템	광주 센터	정보자원 보조플랫폼(안)

※출처: 행정안전부(2019) "공공기관 데이터 자원 보존 개선방안 연구"

본 연구에서는 정보시스템운영은 시스템 인프라 운영차원에서의 데이터 관리에 집중하고 있고, 정보자원보존은 아직 구체적인 실체가 명확하지 않다고 판단하여, 공공데이터 영역에서 진행된 데이터의 품질과 품질관리체계의 내용을 참고하여 기록 관리적인 관점에서의 데이터 세트 품질과 그 평가 방안에 대해서 논의한다. 공공데이터 역시 데이터의 품질이라는 보편적인 관점에서 출발하고 있으며, 기록관리 영역은 그를 참고하되 기록관리라는 독자적인 영역의 관점을 반영하면 충분한

것으로 판단했기 때문이다.

3) 공공데이터의 품질관리

공공데이터는 '개방'에 초점을 맞추어, 데이터의 공개와 활용을 목표로 하고 있다. 그 전제가 되는 것이 공공데이터 품질에 대한 신뢰성을 확보할 수 있는 구체적인 지침과 방안이다. 이를 위해 행정안전부와 한국지능정보사회진흥원(이하 'NIA')은 "공공데이터 예방적 품질관리 진단 가이드"(이하 '품질관리 진단 가이드')를 발간하였다. 품질관리 진단 가이드에서 "고품질의 공공데이터 제공 및 활용을 위해 정보시스템 계획 단계부터 선제적인 품질관리를 시행하여 양질의 데이터 확보 · 관리"하는 것으로 그 목적을 명시하고 있다. 여기서 제목에 포함된 '예방적'과 목적에 서술된 '선제적인'이라는 단어에 주목해야 하는데, 이에 대한 함의는 시스템의 구축 이후의 품질관리는 예산이 많이 소요되며 시스템이 운영 중이므로 개선 활동이 상대적으로 어렵다는 현실적인 사정을 고려하여, 공공데이터의 품질활동은 시스템의 구축 이전에 미리 이루어져야 한다는 것이다. 그에 따려면 정보시스템 구축계획인 ISP/ISMP 단계에서부터 품질관리 활동이 수행되어 최종적으로 시스템 구축요건에 품질요구가 반영되어야 한다.

품질관리의 목적에 이어서 품질관리의 개념은 데이터 표준, 데이터 구조, 데이터값, 데이터 관리체계라는 4개의 진단영역을 관리하는 활동으로 정의되며. 각 진단영역의 대상과 활동은 다음과 같이 제시된다.

- 데이터 표준
  - 데이터 모델 설계 표준, 용어표준, 단어표준, 도메인표준, 코드표준 등에 대한 정의와 함께 지속적인 변화관리 과정을 체계적으로 통제하는 기준과 절차

•데이터 구조

- 데이터 항목들 사이의 배열과 접근 관계를 논리적 관점에서 정의한 것으로 선형 구조, 트리 구조, 네트워크 구조 등으로 표현

•데이터값

- 자료의 객체가 가지고 있는 길이나 형태 따위의 성질을 문자나 숫자로 표현하여 실제로 데이터베이스에 저장하고 있는 값

•데이터 관리체계

- 데이터 품질을 관리하기 위한 목표 설정, 정책 및 조직의 구성, 품질관리 계획의 수립, 표준화, 산출물 관리 및 최신화, 품질 진단 및 개선, 오류신고 관리, 데이터 활용성과 평가, 데이터의 연계 및 개방 등 관리체계 전반에 관한 사항

4개 진단영역은 품질관리 목표와 계획, 조직 구성을 포괄하는 데이터 관리체계를 기반으로 데이터 표준과 구조를 품질관리의 기준으로 설정하고 그 기준에 따라 데이터값들이 유지되고 있는지를 관리하는 것으로 체계화되었다. 이에 따라 다음과 같은 항목에 대해 데이터 품질 진단이 시행된다.

첫째, 데이터 표준 영역으로 데이터 모델 설계 표준, 용어표준, 단어표준, 도메인표준, 코드표준 등에 대한 정의와 함께 지속적인 변화관리 과정을 체계적으로 통제하는 기준과 절차로 구성된다.

둘째, 데이터 구조 영역으로 데이터 항목들 사이의 배열과 접근 관계를 논리적 관점에서 정의한 것으로 선형 구조, 트리 구조, 네트워크 구조 등으로 표현된다.

셋째, 데이터값 영역으로 자료의 객체가 가지고 있는 길이나 형태 따위의 성질을 문자나 숫자로 표현하여 실제로 데이터베이스에 저장하고 있는 값으로 구성된다.

넷째, 데이터 관리체계로 데이터 품질을 관리하는 활동으로 구성된다.

품질관리 활동의 절차는 컨설팅 과정과 실제 구축과정으로 일별 되는데, 전자에서는 4개의 진단영역 전체에 걸친 진단의 결과로 데이터 요구사항이 도출되어 시스템 구축요건에 반영되고, 구축과정에서는 데이터 표준–데이터모델검증–실제 데이터 검증의 절차를 거쳐 시스템 구축요건이 완성되는 것으로 체계화된다.

품질관리 진단 가이드가 공공데이터 품질의 목적과 절차에 대한 원칙을 제시하는 수준이라면, 동일한 기관에서 발간한 “공공데이터 품질관리 매뉴얼”(이하 ‘품질관리 매뉴얼’)에서는 세부적이고 실무에 적용할 수 있는 구체적인 원칙을 제시하고 있다. 품질관리 매뉴얼은 크게 시스템 구축과 운영 과정에서의 품질관리와 데이터 품질 진단과 개선으로 구성되어 있는데, 품질 진단에서는 품질관리 진단 가이드에서 제시한 데이터 값, 데이터 구조, 데이터 표준을 각각 정확성, 완전성, 일관성이라는 3대 품질지표에 대응시키고 각각의 지표에 대해서 전체 데이터에 대한 오류데이터의 비율(오류율) 측정을 하는 것을 핵심 방법론으로 삼고 있다.

오류율 산정 기준은 구체적인 계산식으로 제시된다. 3대 품질지표별 오류율에 요소별 가중치를 적용한 값의 합계로 개별 행정정보데이터세트의 전체 품질 오류율을 산출하게 되고, 개별 지표별 오류율은 진단항목별로 평가한 데이터의 총 건수에 대한 오류데이터의 비율로 산출된다. 지표별 품질 오류율과 품질 요소별 오류율 계산식은 다음과 같다.

$$\text{품질오류율}(\%) = \sum_{i=1}^{n}(Ei \times Wi)$$

$$\text{값 오류율}(\%) = (\sum_{i=1}^{n} ei \div \sum_{i=1}^{n} si) \times 100$$

품질 오류율 (E=품질 요소별 오류율, W= 품질 요소별 가중치)
값 오류율 (i=진단항목, s = 전체 데이터 건수, e = 오류데이터 건수)

전반적인 품질관리 수준의 현황을 파악하기 위해 위에 서술된 주요 품질지표 3개(정확성, 완전성, 일관성)에 덧붙여 7개의 지표가 제시되는데, 그 결과는 〈표 2〉과 같다

〈표 2〉 공공데이터 품질지표

지표	세부지표	내용
준비성	- 관리지표 - 내용 충실	공공데이터의 품질관리를 위해 기본적으로 관리해야 하는 정책, 규정, 조직, 절차 등을 마련하고, 최신의 내용으로 충실하게 관리되는지를 측정하는 지표
완전성	- 논리모델(논리적 데이터모델링) - 식별자 - 물리 구조(물리적 데이터모델링) - 속성의미	공공데이터의 저장소인 데이터베이스를 구축함에 있어 논리적인 설계와 물리적인 구조를 갖추고, 업무요건에 맞게 데이터가 저장되는지를 측정 하는 지표
일관성	- 속성(attribute) - 표준 - 중복 값 - 연계 값	같은 의미를 갖는 데이터는 논리적 속성 단위, 물리적 컬럼 단위에서 일관된 이름과 형식을 갖도록 표준을 준수하고 있는지, 공공 데이터의 공동 활용을 위해 공유·연계하는 데이터는 누락이 없이 상호간의 일관성을 유지하는지를 측정하는 지표
정확성	- 입력값 - 업무규칙(BR, Business Rule) - 범위·형식 - 참조 관계 - 계산식	정확한 데이터 제공을 위해 데이터의 입력 단계부터 오류가 입력되지 않도록 하고, 저장된 데이터가 정의된 기준에 맞게 유효한 값의 범위와 형식으로 되어 있는지, 저장된 데이터가 현실에 가장 가까운 최신 값을 반영하고 있는지를 측정하는 지표
보안성	- 데이터 소유권, 관리권 - 접근제한 - DB 보호	정확한 데이터 제공을 위해 데이터의 입력 단계부터 오류가 입력되지 않도록 하고, 저장된 데이터가 정의된 기준에 맞게 유효한 값의 범위와 형식으로 되어 있는지, 저장된 데이터가 현실에 가장 가까운 최신 값을 반영하고 있는지를 측정하는 지표
적시성	- 응답시간 - 데이터제공 - 최신값	사용자가 만족하는 수준의 응답시간이 확보 되고 잇는지, 사용자의 데이터 요구에 따른 수집·처리·제공까지의 절차가 체계적으로 관리되고 있는지를 측정하는 지표
유용성	- 충분 - 접근 - 활용	사용자가 만족하는 수준의 충분한 정보가 제공되고 있는지, 정보 접근 시 사용자의 편의성이 확보되고 있는지, 사용자의 정보 이용에 따른 만족 수준을 높이도록 노력하고 있는지를 측정하는 지표

※출처: 한국정보화진흥원 “품질관리 매뉴얼 표 II-8. 품질지표“에서 발췌

이러한 지표를 사용하여 품질 진단이 진행되며, 데이터의 형태와 특성에 따라 다양한 방법을 사용하여 진단을 수행한다. 품질 진단의 방법은 다음과 같다.

첫째, 프로파일링을 통한 값과 구조에 대한 기술적인 분석방법이다. 컬럼, 날짜, 코드, 참조무결성을 대상으로 하는데, 컬럼에서 시작하여 단일 테이블, 테이블 간 연계 관계 등 그 대상을 단일 값에서 데이터 구조 수준으로 확대하는 방법으로 진행하게 된다. 예를 들어, 컬럼 수준의 프로파일링을 통해 접수 테이블 컬럼의 접수 일자는 8자리(YYYYMMDD)의 유효한 날짜형식을 유지해야 한다는 규칙을 검증한다면, 단일 테이블 수준에서는 접수 여부가 'Y'이면 접수 일자는 반드시 8자리의 유효한 날짜형식의 데이터를 유지해야 한다는 규칙을 검증하는 식이다.

둘째 인터뷰 설문을 통한 방법이다. 이는 인터뷰와 설문을 통해 전반적인 데이터 품질관리 수준과 지표별 데이터 품질 수준을 체크리스트를 통해 진단하는 방법이다.

셋째, 업무규칙 진단이다. 이는 법, 규정에 정의된 업무 기준(계산식)에 근거하여 데이터가 관리되고 있는지를 진단하는 방법으로서, 업무규칙을 준수하고 있는지에 관한 측정 스크립트(SQL 등)를 실행하여 오류 값을 추출하는 방법이다.

넷째, 비정형 실측이다. 이는 문서, 이미지, 동영상 등 정형화되어 있지 않은 정보를 사람이 직접 확인(실측)을 통하여 오류 여부를 진단하는 방법으로써, 별도 도구 없이 직접 정보를 조회하거나 해당 문서를 수기로 확인하는 방법이다.

그 결과 공공데이터의 품질관리 결과는 최종적으로 5등급으로 구성되어 〈표 3〉과 같은 평가등급이 책정된다.

〈표 3〉 공공데이터 품질관리 평가등급

등급	설명
5레벨	조직 전체의 데이터 품질관리 활동의 선순환 체계가 확립되고, 이를 통해 공공데이터의 안정적 품질 향상 및 유지가 보장되는 수준
4레벨	조직 전체의 데이터 품질관리 프로세스가 이행되고, 데이터 품질관리 활동 수행에 따른 성과측정이 가능한 수준
3레벨	데이터 품질관리를 위한 전반적인 활동들이 관리 및 통제되어, 이를 통해 데이터 품질 향상이 가능한 수준
2레벨	데이터 품질관리가 인식되고, 품질 진단에 따른 개선 조치 등 기본적인 품질관리 활동들을 수행하는 수준
1레벨	데이터 품질관리가 인식이 미흡하여 기본적인 품질관리 활동의 수행이 불가능하거나 부분적인 품질관리 활동만 수행되는 수준

※출처: 한국정보화진흥원 “품질관리 매뉴얼 II.데이터 품질관리 2.기관의 데이터 품질관리” 에서 발췌.

4) 데이터세트 데이터 품질평가의 설계 방향

데이터 품질평가의 기본적인 방법론은 정보기술 보안관리를 위한 국제표준 지침인 ISO/IEC 13335-1의 위험분석 전략의 베이스라인 접근법을 준용하였다. 베이스라인 접근법은 대상에 대해서 표준화된 보호 대책의 세트를 체크리스트 형태로 제공하는 것으로서, 체크리스트에 있는 보호 대책이 현재 구현되어 있는지를 조사하여 구현되지 않은 보호대책을 식별하는 방법이다. 베이스라인 접근법 외에 비정형 접근법, 상세 위험분석, 복합 접근법이 있으나, 아직 데이터세트 평가 방법에 대한 구체적인 결과와 피드백이 이루어지지 않은 상태에서 평가가 과중한 업무 부담이 될 수 있는 것을 고려하여 가장 기초적인 베이스라인 접근법을 참고하였다.

이를 기반으로 데이터세트의 데이터 품질평가는 공공데이터 품질관리를 참고하되, 무결성이라는 기록관리 대원칙을 보장한다는 관점에서 현실적으로 유용하게 평가될 수 있도록 설계하는 것을 원칙으로 하였

다. 공공데이터 품질관리와의 내용적 차이와 그에 따른 설계 방향은 다음과 같다.

첫째, 공공데이터 품질관리가 예방적이고 선도적인 차원에서의 데이터를 '관리'하는 목적이라면, 데이터세트 품질평가의 목적은 데이터세트 기록의 '평가'의 일환으로서 데이터세트 식별 이후 내용적 가치, 데이터 연계 가치에 대한 가치평가가 이루어진 데이터세트에 대해서 최종적으로 품질을 확인하게 된다.

〈표 4〉 행정정보데이터세트의 평가선별 절차

STEP 1	데이터세트 식별	시스템 구조식별
		시스템 세부식별
⇓		
STEP 2	업무기능 및 내용 가치평가	업무 활용가치 평가
		증빙 가치평가
		역사적·학술 가치평가
⇓		
STEP 3	데이터 연계·결합가치 평가	통계 활용성
		내부자원 활용성
		외부자원 활용성
⇓		
STEP 4	데이터 품질평가	품질관리 활동
		메타데이터 및 데이터 품질
		디지털객체 품질
⇓		
STEP 5	기록관 보유 기간 책정	
⇓		
STEP 6	이관 및 수집 결정	

※출처: 국가기록원(2021년) "행정정보데이터세트 기록정보 서비스 및 활용모형 연구 최종보고서"에서 발췌.

〈표 4〉는 2021년 연구에서 제시된 행정정보데이터세트의 평가선별 절차이다. 데이터세트 품질 평가는 가치 판단의 영역이라기보다는 최

종적인 품질관리상태를 확인하고 그 결과에 따라 생산기관 계속 보유 여부, 혹은 이관 여부 등을 판단하는 객관적인 기준을 제시하는 것을 목적으로 삼았다. 데이터 품질평가를 통해 데이터 품질이 지속적으로 하락할 수 있다고 판단이 되면 생산시스템의 품질요건을 강화하거나 품질관리 활동을 권고할 수 있으며, 데이터 품질에 심각한 문제가 있을 경우, 영구기록물관리기관으로 즉각적인 이관을 결정할 수 있다(STEP 6).

둘째, 공공데이터 품질관리에서 품질 진단이 시스템의 모든 데이터를 대상으로 한다면, 데이터세트 데이터 품질평가는 특정 행정정보시스템의 데이터 전체의 부분집합인 데이터세트를 대상으로 한다. 행정정보데이터세트는 업무적으로 연계성이 있는 데이터의 집합(세트)을 논리적으로 구성한 결과로서 그 대상은 전체 데이터의 부분집합일 수밖에 없기 때문이다.

셋째, 공공데이터 품질관리에서는 주로 관계형데이터베이스을 대상으로 값과 구조를 진단하는데 초점을 맞추고 있다면 데이터세트 데이터 품질평가는 평가의 대상에 첨부파일로서 디지털객체를 포함했다. 이는 기록관리 관점에서는 관계형 데이터베이스에서 관리되고 있는 기록의 메타데이터와 값뿐만 아니라, 업무 행위의 실체로서의 디지털객체인 첨부파일에 대한 품질평가가 필요하기 때문이다.

이러한 관점에서 데이터세트의 품질평가는 공공데이터 품질관리의 지표를 참고하여 3개의 영역으로 설계되었다. 데이터 표준과 데이터 관리체계를 합쳐 품질관리 활동 영역, 데이터 구조와 값 영역은 메타데이터 및 데이터 품질 영역으로 설정하고, 기록의 첨부파일을 대상으로 하는 디지털객체 품질 영역을 추가하였다. 데이터 표준과 관리체계를 합친 이유는 기록관리의 원칙인 무결성의 관점에서 보았을 때, 데이터세트의 데이터 품질은 시스템 구축 이후의 업무활동에서 이루어지는 기록 생산 과정의 맥락에서 무결성이 보장되는 것으로 충분하다고 판

단했고, 실질적으로 품질평가 활동이 공공데이터의 영역과 기록관리의 영역에서 중복되어 이루어질 필요는 없다는 현실적인 고려사항을 반영한 것이다. 이렇게 설계된 데이터세트의 데이터품질관리 영역 및 기본지표와 공공데이터의 품질관리 영역을 비교한 결과는 최종적으로 〈표 5〉와 같다.

〈표 5〉 공공데이터 품질관리 영역 및 기준과 데이터세트 데이터품질관리 영역 비교

공공데이터 품질관리		데이터세트 데이터품질관리
데이터관리체계	준비성	품질 관리활동 영역
데이터 구조	완전성	메타데이터 및 데이터 품질영역
데이터 표준	일관성	메타데이터 및 데이터 품질영역
데이터값	정확성	메타데이터 및 데이터 품질영역
데이터관리체계	보안성	품질 관리활동 영역
데이터관리체계	적시성	메타데이터 및 데이터 품질영역
데이터관리체계	유용성	메타데이터 및 데이터 품질영역
해당 사항 없음		디지털객체 품질영역

3. 데이터세트 데이터 품질평가 지표개발

1) 품질관리 활동 영역

품질관리 활동은 데이터세트 데이터에 대한 품질활동이 기관의 공식적인 업무활동으로 이루어지고 있는지를 평가하는 것이다. 구체적으로는 데이터 품질에 관련한 목표와 정책 설정, 업무와 조직 구성, 업무협조체계, 품질관리 계획의 수립, 개발 및 운영 표준, 설계 등의 관련 문서의 확보 및 최신화 현황, 품질 진단 및 오류 사항 관리, 개선 활동 등으로서, 품질과 관련된 기관의 활동 전반을 포함한다. 여기서 주안점을 두어야 하는 것은 품질관리 활동이 기관의 정규 업무로서 주기적으로

이루어지고 있으며, 결과에 대한 개선과 대응이 다음 평가주기의 활동에 반영되고 있는가이다. 즉, 품질관리 활동이 단발성으로 끝나지 않고 활동 결과와 개선이 선순환 구조를 이루고 있는지를 평가의 핵심으로 삼아야 한다는 점이다. 이를 위해서 다음과 같은 기준이 필요하다.

첫째, 품질관리와 그 활동에 대해서 기관 차원에서 공식적으로 문서화된 지침의 존재 여부와 내용의 충실성이 평가되어야 하고, 지침에 따라 품질관리 활동이 수행되는가에 대한 평가가 이루어져야 한다. 그리고 품질관리지침의 내용은 품질관리의 구체적인 목표와 방향, 수행 조직, 평가 절차, 진단계획, 평가 결과에 따른 개선계획이 포함되어야 한다. 특정한 품질관리지침 없이 품질관리 활동을 수행하고 있을 경우가 있을 수 있는데, 이럴 때는 해당 행정정보시스템의 구축과 운영 과정에서 만들어졌던 산출물(데이터 사전 정의서, 도메인 정의서, 코드 정의서, 엔티티 정의서, ERD, 테이블 정의서, 컬럼 정의서, DB 정의서, 업무규칙 정의서 등)을 통해 간접적으로 품질관리 활동의 지침으로 평가할 수 있다. 단, 향후 관련 내용을 반영한 품질관리 지침의 생산을 권고해야 한다.

둘째, 데이터의 기본적인 신뢰성이 확보되었는지의 여부이다. 여기에서 신뢰성은 데이터의 값에 대한 신뢰성(유효성)이 아니라, 품질관리 대상이 되는 데이터의 양이 충분한지, 향후 데이터에 대한 추가 변경이 없을 것이라는 신뢰성을 의미한다. 예를 들자면, 특정 컬럼이 널값(null value)을 허용할 경우 해당 컬럼에 데이터값이 저장된 레코드(row)의 숫자가 표본으로 삼을 정도로 충분한지를 확인하는 것이다. 해당 데이터의 행 숫자가 전체 행 숫자에 비교하여 너무 적으면, 데이터의 결측으로 볼 수는 없으나, 해당, 컬럼에 대한 데이터베이스 정규화를 권고할 수 있다. 그리고 업무규칙에 따라 데이터가 종결(업무 마감 등)되었다면, 앞으로 해당 데이터에 대한 추가나 변경, 삭제가 있을 수 없다는 점이 보장되어야 한다.

이에 따라 품질관리 활동 영역에서는 품질관리지침 및 활동 여부, 원자료(raw data)의 신뢰성이라는 평가지표가 도출된다.

2) 메타데이터 및 데이터 품질 영역

메타데이터 및 데이터 품질평가는 데이터세트의 개체 무결성(entity integrity), 참조 무결성(referential integrity), 도메인 무결성(domain integrity)에 해당하는 영역으로 구조와 값의 정합성을 평가하는 것이다. 구체적으로 데이터의 구조와 값의 변경이 별도의 이력 정보로 관리되고 있는지, 데이터 간의 참조 관계가 정확한지, 메타데이터의 명칭과 데이터값에 대한 규칙이 명확한지 등에 대한 평가이다. 여기서 주안점을 두어야 하는 것은 데이터 수준에서의 개별적인 진단도 중요하지만, 구조와 내용에 대한 규칙이 해당 데이터 내부가 아니라 외부에 독립적으로 존재하느냐에 중점을 두고 평가해야 한다는 점이다. 이는 데이터세트의 평가는 향후 지속적으로 이루어질 것이며, 이에 따라 보유기간이 길어지거나 이관이 되면서 평가의 주체가 바뀔 수 있으며, 그 시점의 평가자가 품질을 판단하는 데 추측이나 유추가 아니라 명시된 규칙에 따라 판단할 수 있어야 하기 때문이다. 이를 위해서 다음과 같은 기준이 필요하다.

첫째, 데이터 변경 이력관리 여부이다. 데이터값의 변경 이력을 저장하는 별도의 이력 정보가 존재해야 하며, 이를 기초로 이력 정보의 정확성을 평가할 수 있어야 한다. 변경 이력의 관리는 데이터의 증빙적 가치와 보안성 보장의 근거가 된다.

둘째, 데이터베이스의 구조와 참조무결성이다. 컬럼의 추가, 컬럼 값의 제약조건(constraint), 테이블 간 연계 관계(relation) 업데이트 등 데이터 구조의 변경이 이루어질 경우, 그 내용이 적절히 기록되고 현행화되고 있는가를 평가할 수 있어야 한다. 참조 관계가 명확하지 않을 경우,

데이터 신뢰성에 문제가 발생할 가능성이 커지게 되므로 주키(Primary Key)와 외래키(Foreign Key)를 통해 정의되는 테이블 간 참조 관계가 명확한지, 정상적으로 연계되고 있는지 등을 파악할 수 있어야 한다.

셋째, 메타데이터의 기술(description)정보 유무이다. 컬럼의 속성 등 메타데이터의 의미정보와 테이블의 명칭, 사용자 접근 권한 등이 정확하게 부여되어 관리되고 있고, 별도의 문서 등으로 관리되고 있는가를 평가할 수 있어야 한다. 데이터세트가 저장된 데이터관리시스템에서 이러한 기술정보를 별도로 저장할 수 있는 기능을 제공할 경우 해당 기능의 적절한 사용 여부를 평가할 수 있다. 〈그림 2〉에서는 오라클 사의 시스템 뷰 테이블에서 데이터베이스 컬럼에 대하여 설계 단계에서 기재한 주석(comment)을 조회한 결과를 보여주고 있다.

〈그림 2〉 Oracle의 comment view

HR_ORCL | EMPLOYEES

Columns | Data | Constraints | Grants | Statistics | Column Statistics | Triggers | Dependencies | Details | Partitions | Indexes | SQL

Actions...

Column Name	Data Type	Nullable	Data Default	COLUMN ID	Primary Key	COMMENTS
EMPLOYEE_ID	NUMBER(6,0)	No	(null)	1	1	Primary key of employe...
FIRST_NAME	VARCHAR2(20 BYTE)	Yes	(null)	2	(null)	First name of the emplo...
LAST_NAME	VARCHAR2(25 BYTE)	No	(null)	3	(null)	Last name of the emplo...
EMAIL	VARCHAR2(25 BYTE)	No	(null)	4	(null)	Email id of the employee
PHONE_NUMBER	VARCHAR2(20 BYTE)	Yes	(null)	5	(null)	Phone number of the e...
HIRE_DATE	DATE	No	(null)	6	(null)	Date when the employe...
JOB_ID	VARCHAR2(10 BYTE)	No	(null)	7	(null)	Current job of the emplo...
SALARY	NUMBER(8,2)	Yes	(null)	8	(null)	Monthly salary of the e...
COMMISSION_PCT	NUMBER(2,2)	Yes	(null)	9	(null)	Commission percentage...
MANAGER_ID	NUMBER(6,0)	Yes	(null)	10	(null)	Manager id of the emplo...
DEPARTMENT_ID	NUMBER(4,0)	Yes	(null)	11	(null)	Department id where e...

※출처: SQL Developer for Database Developers An Oracle White Paper에서 발췌.

넷째, 유효성 규칙 정의 여부이다. 법·제도 및 규정에 따라 진행되는 업무의 규칙을 반영하는 데이터 값의 유효성 규칙이 업무규칙 정의

서 등 별도의 문서로 정의되어 있는지를 확인하고, 그 내용이 데이터베이스 설계에 반영되어 있는지를 판단한다. 그리고 데이터 누락 여부, 조건, 범위 등 데이터값의 유효성 규칙이 ERD, 컬럼 정의서 등 별도의 문서에 정의되어 있는지를 확인하고 해당 규칙을 준수하는가를 평가할 수 있어야 한다. 〈표 6〉은 업무규칙과 관련 근거를 고려하여 품질 평가를 수행하는 방법에 대한 예시이며, 〈표 7〉은 데이터값의 유효성 규칙이 명시적으로 정의된 예시이다.

〈표 6〉 업무규칙의 반영 예시

제도 및 규정 근거	업무규칙	평가 방법
국가회계법 국가회계법 시행령 제3조	당해에 회계감사를 의뢰할 회계법인을 선정해야 하며, 특정 공인회계사가 3년 이상 연속으로 회계감사를 계속 수행할 수 없다.	감사 테이블–감사인 컬럼–감사 연도 데이터의 값을 진단

〈표 7〉 유효성 규칙 예시

데이터 유효성 규칙	관련 데이터값
필수 컬럼에는 값의 누락이 없어야 한다.	생산자, 사용자 아이디 등
조건에 따라 컬럼 값이 항상 존재해야 한다.	지리정보의 경위도 등
컬럼 값이 해당 정보의 유형에 따라 유효한 값을 가져야 한다.	세계표준시 사용 시 기준 시간 표기 여부, 주민등록번호 자릿수 13자리 및 주민등록번호 생성 규칙
복수의 컬럼 값이 선후 관계에 있으면 이 규칙을 지켜야 한다.	중간결산 일자 이후 최종결산 일자 저장
정보의 발생, 수집, 그리고 갱신 주기를 유지해야 한다.	최근의 레코드(row) 정보와 이력 정보 일치

다섯째, 코드 정의 표준성이다. 데이터값이 특정한 속성에 따라 그룹으로 분류되어 있을 경우, 데이터 조회의 신속성이나 데이터베이스 정규화를 위하여 데이터를 코드화할 경우가 있는데, 코드화의 원칙 없이 자의적으로 코드를 설정하고 그 설정 원칙을 별도로 정의해두지 않을

경우, 해당 데이터세트에 대한 활용성이 급락하게 된다. 이를 방지하기 위하여 공신력 있는 코드를 준용하는지에 대한 여부와 자체적인 코드 사용 시 그 원칙이 외부에 별도로 정의되어 있는지를 평가한다. 공신력의 여부는 〈표 8〉에 예시로 든 행정안전부의 행정정보코드처럼 코드가 표준화되어 온라인 등으로 게시되었는지에 따라 판단할 수 있다.

〈표 8〉 코드 정의 예시

코드	예시
행정안전부 행정정보코드	국가 자격 면허목록 0110001: 기계기술사
식품의약안전처 의약품 표준코드	A주 바이알 10ml 1개: 8800001123411
금융투자협회 협회 표준코드	메리츠 W 사모 채권 투자신탁 3호: KR5501730455

이에 따라 메타데이터와 데이터 품질 영역에서는 데이터 변경 이력 관리, 데이터베이스의 구조와 참조무결성, 메타데이터 기술정보, 유효성 규칙 정의, 코드 정의 표준성이라는 평가지표가 도출된다.

3) 디지털객체 품질 영역

디지털객체 품질 영역은 기록에 첨부된 첨부파일로서의 디지털객체를 대상으로 그 객체에 대한 접근경로의 유효성과 객체 자체의 재현성을 평가하는 영역이다. 이 영역은 기록 관리적인 관점, 특히 공공기록관리체계에서의 기록관리 관점에서 중요하다. 공공기록관리체계에서는 공공기록이 업무활동의 결과라는 대원칙을 가지고 있는데, 기록물의 메타데이터를 통해 업무활동의 맥락과 성격을 파악할 수 있다면 디지털객체는 기록물건에 첨부되는 첨부파일로서 업무활동 자체의 직접적인 결과이기 때문이다. 디지털객체의 품질을 평가하기 위해 다음과 같은 기준과 지표가 필요하다.

첫째, 유형별 생산지침이다. 데이터베이스가 참조하고 있는 디지털 객체의 유형에 따른 생산지침 유무와 준수 정도를 평가한다. 생산지침의 예를 든다면, 문서 작성기로 작성된 문서파일의 경우 표준포맷(*.~x)으로의 저장 여부 등이 지침으로 명시될 수 있고, 〈표 9〉처럼 포맷별 기술항목이 지침으로 명시될 수 있다. 기관 수준에서 생산지침을 가지고 있지 않다면, 국가기록원에서 제시하고 있는 관련 지침을 준용할 수 있다.

〈표 9〉 포맷별 요소의 상호참조표

	통합서지표준	유물분류표준	작품관리표준
기술	256 컴퓨터파일의 특성 355 보안분류 506 접근제한주기 516 C파일/데이터유형주기 538 시스템사항주기 565 파일특성주기 753 컴퓨터파일접근시스템 754 C파일기계분류식		검색엔진번호(2)

※출처: 조윤희(2003) "문화콘텐츠 통합을 위한 메타데이터 포맷 연구(II)"에서 발췌.

둘째, 경로 유효성이다. 디지털객체가 행정정보시스템의 특정 디렉터리 경로에 저장되고 데이터세트의 메타데이터에서 디지털객체의 경로 정보를 저장하고 있을 때, 데이터세트가 참조하고 있는 디지털 컴포넌트의 경로 유효 여부를 평가한다. 디지털객체가 경로 정보를 가지고 있지 않고 BLOB(Binary Large Object) 형태로 데이터베이스에 바로 통합되어 있다면, 해당 객체의 바이너리 데이터가 유효성 진단의 대상이 된다. 경로 유효성은 특히 데이터세트의 활용과 이관에 대비하여 반드시 평가되어야 할 지표이다.

셋째, 사용 가능성이다. 디지털객체를 실행하여 생산 당시의 화면 렌

더링과 룩앤필(Look & Feel)을 재현할 수 있는 응용프로그램이 평가 시점에서의 단종, 기술지원 여부 등을 평가한다. 사용 가능성은 데이터세트의 장기보존과 관련해서 평가되어야 할 지표인데 이를 위해서 영국의 TNA에서는 PRONOM이라는 디지털객체 기술정보서비스를 통해 디지털객체의 파일포맷 사용 가능성을 제시하고 있으며(〈그림 3〉), 국가기록원에서도 기술정보은행(Digital Format Registry, DFR)을 통해 파일포맷 기술정보를 제공하고 있다.

〈그림 3〉 PRONOM의 Microsoft Word 2007버전 기술정보

Details for: Microsoft Word for Windows 2007 onwards

Save as... XML | CSV Print

Go to: Summary | Documentation > | Signatures > | Compression > | Character encoding > | Rights > | Reference files > Properties >

Summary

Name	Microsoft Word for Windows
Version	2007 onwards
Other names	
Identifiers	MIME: application/vnd.openxmlformats-officedocument.wordprocessingml.document PUID: fmt/412
Family	
Classification	Word Processor
Disclosure	
Description	From Microsoft Office 2007 onwards, the core output format of MS Word has been based on the Office Open XML (OOXML) file format. The ISO standard for OOXML is ISO/IEC DIS 29500. An OOXML file format consists of a compressed zip archive that is designated according to which file type it is. Further detail on OOXML can be found within fmt/189 - Microsoft Office Open XML. An alternative extension of .wbk refers to a backup file of a Word document, however there is no material or structural difference between a .wbk file and the .doc file it is a backup of.
Orientation	
Byte order	
Related file formats	Is supertype of Microsoft Word Template (2007 onwards) Is supertype of Microsoft Word Macro-Enabled Document Template (2007 onwards)

※출처: TNA 홈페이지에서 발췌.

넷째, 공개 표준성이다. 이를 통해 디지털객체의 공개표준 여부를 평가한다. 예를 들어 문서파일 데이터의 경우 개방형 문서형식(Open Document Format for Office Applications, ODF)인 odt 형식으로 저장되었

는지를 평가한다. 공개 표준성은 데이터세트의 장기보존과 관련해서 평가되어야 할 기준이며, 특히 인공지능의 활용을 위한 기계 가독성(machine readablity)을 확보할 수 있는 핵심 지표이다.

4. 국가철도공단의 데이터세트 데이터 품질평가 적용사례

1) 데이터세트 품질평가 도구 설계

앞서 제시한 데이터세트의 데이터 품질평가 기준에 따라 평가도구의 기본 골격을 설계하였다. 평가도구의 설계 방향은 위험분석 전략의 베이스라인 접근법을 참고하였으며 구성은 영역별 지표에 따른 구체적인 질문과 담당자, 체크리스트의 형태의 답변(〈표 10〉)으로 구성되었다.

평가의 답변은 업무규칙을 잘 이해하고 있는 처리과와 시스템을 관리하는 시스템담당자가 가능하면 공동으로 진행하도록 구성되어 있다. 답변 결과는 예/아니요로 구성된 체크박스를 선택하고 답변 결과에 따라 발생 가능성을 상/중/하에서 선택하게 구성되어 있다. 이는 답변이 서술형으로 이루어질 경우 답변자의 주관이 개입되거나 각 지표에 대한 균등한 수준의 답변이 이루어지지 않을 가능성을 고려한 것이며. 답변의 객관적인 기준은 개별 기관의 상황에 맞추어 설정할 수 있다. 예를 들어 데이터 보유량이 방대하지만, 시스템 운영단계에서 데이터품질관리 활동이 활발하게 이루어지고 있는 기관은 표본 평가를 선택하고, 오류에 대한 평가 기준을 엄격하게 잡을 수 있다. 그리고 품질관리의 기준이 엄격한 기관에서는 데이터값의 유효성 규칙을 위반한 경우가 20%를 넘을 경우, 그 답변을 '아니요'로 판정하고 발생 가능성이 크다고 평가할 수 있다.

〈표 10〉 데이터세트 품질평가 도구

영역	질문	담당자	예/아니오	발생 가능성
품질관리 활동	데이터품질관리 지침이 부재하여 데이터 품질관리 활동에 문제가 발생할 가능성이 있습니까?	처리과 시스템 담당자	□예 □아니요	□상 □중 □하
	데이터 품질관리 활동이 부재하여 데이터의 품질에 문제가 발생할 가능성이 있습니까?	처리과 시스템 담당자	□예 □아니요	□상 □중 □하
	분석에 필요한 원 데이터가 양이 부족할 가능성이 있습니까?	시스템 담당자	□예 □아니요	□상 □중 □하
	원 데이터의 신뢰성이 부족할 가능성이 있습니까?	시스템 담당자	□예 □아니요	□상 □중 □하
	오류데이터 발견 시 적절한 데이터 보정이 이루어지지 않을 가능성이 있습니까?	처리과 시스템 담당자	□예 □아니요	□상 □중 □하
메타데이터 품질	데이터값 변경 시 이력 정보를 생성하고 있지 않아 데이터 변경 이력을 추적하지 못할 가능성이 있습니까?	처리과 시스템 담당자	□예 □아니요	□상 □중 □하
	시스템 운영 중 데이터세트 항목의 추가, 데이터값 구조 변경 등의 변화가 있습니까?	처리과 시스템 담당자	□예 □아니요	□상 □중 □하
	테이블 간 연결정보(주키와 연결키)가 데이터베이스에 정의되지 않는 경우가 있습니까?	시스템 담당자	□예 □아니요	□상 □중 □하
	테이블 명과 컬럼 명의 의미정보가 데이터베이스에 정의되지 않는 경우가 있습니까?	시스템 담당자	□예 □아니요	□상 □중 □하
	데이터값 유효성 규칙이 데이터베이스에 정의되지 않는 경우가 있습니까?	시스템 담당자	□예 □아니요	□상 □중 □하
	사용되는 코드의 의미정보가 데이터베이스에 정의되지 않는 경우가 있습니까?	시스템 담당자	□예 □아니요	□상 □중 □하
디지털객체 품질	데이터베이스가 참조하고 있는 디지털 컴포넌트에서 해당 유형에 따른 생산지침이 없는 경우가 있습니까?	처리과 시스템 담당자	□예 □아니요	□상 □중 □하
	데이터베이스가 참조하고 있는 디지털 컴포넌트의 경로가 유효하지 않은 경우가 있습니까?	시스템 담당자	□예 □아니요	□상 □중 □하
	디지털 컴포넌트를 실행할 수 있는 응용프로그램이 단종되거나 기술지원이 중지되어 디지털 컴포넌트를 실행할 수 없는 경우가 있습니까?	처리과 시스템 담당자	□예 □아니요	□상 □중 □하
	디지털 컴포넌트의 파일포맷이 공개표준(odt 등)을 지원하지 않는 경우가 있습니까?	처리과 시스템 담당자	□예 □아니요	□상 □중 □하

평가도구는 어디까지나 일종의 평가 준거틀(프레임워크)로 제시되었으며, 기관의 상황에 따라 평가지표와 질문에 관한 결과를 정성적으로 표기해야 할 경우 별도의 항목에 기재 할 수 있으며 서식과 표기방법의 변경도 가능하다.

다음은 2020년 상반기에 KR에서 진행된 구체적인 데이터품질 평가의 사례이다. 이를 통해 데이터품질평가가 데이터세트 평가의 한 과정으로서 어떻게 이루어지는지에 대한 대략적인 구조와 절차를 인지할 수 있고, 품질평가 지표와 평가도구의 실제 적용사례를 살펴볼 수 있을 것으로 판단된다.

2) KR의 데이터세트 평가 원칙과 절차

KR의 평가 원칙은 2021년 연구에서 제시한 데이터세트 평가 절차와 지표를 준용하였다. 첫째, 현재 데이터의 품질관리 활동, 메타데이터 품질 및 디지털 개체 품질을 확인하여 평가 결과에 따라 데이터 정제, 이관 시점과 여부를 판단하였다. 둘째, 데이터 품질이 하(下)일 경우, 증빙적·활용적 목적으로 데이터세트 기록을 이관해야 할 때 데이터 정제 후 품질을 재평가 하고 이관하도록 였다. 셋째, 행정정보시스템의 운영환경이 변화함에 따라 데이터세트 기록 품질은 주기적으로 확인하여 중요기록의 이관 여부를 판단하도록 하였다.

평가 절차 역시, 앞서 2021년 연구에서 제시한 평가 절차를 준용했다. 시스템 분석단계에서 관리대상의 선정기준을 업무상의 중요성과 통계나 데이터가공시스템의 기반데이터 보유 여부로 잡고, 행정정보시스템 46개 중 사업관리시스템 2종, 전자조달시스템 1종, 재산관리시스템을 데이터세트로 관리대상으로 선정하였다. 선정된 데이터세트의 식별을 수행하였고, 사업관리시스템 데이터세트 기록 29개(고유 26개, 공통 3개), 전자조달시스템 데이터세트 기록 8개(고유 4개, 공통 3개), 재

산관리시스템 데이터세트 기록 16개(고유 13개, 공통 3개)를 식별할 수 있었다.

식별을 통해 파악된 데이트세트에 대하여 평가를 진행하였고, 평가 선별 도구를 업무기능 영역, 내용 가치 영역, 데이터 활용 및 품질 가치 영역으로 구분하여 총 23개의 평가지표에 따라 데이터세트 기록별로 평가했다. 업무기능 및 내용 가치 영역에서 파악한 보존기간(단위업무 등)을 참고하여 보존기간을 1차로 책정하고, 활용 및 품질 가치 영역 평가 결과에 따라 보존기간의 상향 등을 고려하였다.

평가 결과를 바탕으로 기록물관리 전문요원이 데이터세트 별로 보존기간을 최종적으로 책정하고 이관 또는 수집대상을 선별하게 된다, KR의 경우, 보존기간 책정 기준을 기존의 보존기간 7종을 사용하지 않고 업무 활용과 내외부 데이터와의 연계결합 가능성을 바탕으로 자유롭게 책정하였고, 이관 시점과 주기는 업무활동과 데이터 운영 및 이관 비용을 고려하여 결정하였다. 그 결과는 데이트세트 평가심의위원회에 제출되어 최종적으로 승인된 후, KR은 자체관리기관으로 기록관에서 영구기록물관리를 위해 자체개발한 ARMS(영구기록물관리시스템)을 운영하고 있으므로 선별된 이관 및 수집대상 데이터세트 기록은 ARMS으로 이관될 예정이다.

3) KR의 데이터 품질 평가도구 적용 결과 분석

품질관리 활동 영역에서는 각 시스템 전체를 대상으로 한 품질관리 지침은 없으나, 품질관리활동은 시스템의 개별 업무별로 관련 지침과 모니터링 활동이 존재하는 것으로 확인되었다. 원 데이터 신뢰성의 경우 데이터의 양 자체는 모든 데이터세트가 충분하다고 평가되었으나, 데이터의 질에 대해서는 시스템별로 상이한 평가 결과가 나왔다. 분석 결과, 전자조달시스템을 제외한 시스템의 경우 전자결재 등 내부 시스

템연계 등으로 데이터의 질이 기본적으로 보장될 수 있으나, 전자조달 시스템의 경우 조달청의 나라장터 시스템의 공고정보가 수작업을 통해 입력되었으므로, 공고 후 유찰이나 지명입찰 등으로 공고정보가 변경되었을 때 적시에 동기화되지 않아 전자조달시스템의 데이터 수가 맞지 않는 경우가 있었기 때문으로 분석되었다. 이를 개선하기 위하여 나라장터 시스템의 API 연계 가능성을 확인하고, 가능하면 연계할 수 있도록 권고하였다.

메타데이터와 데이터 품질 영역에서는 데이터 이력관리와 구조, 의미정보에서는 전반적으로 문제가 없는 것으로 평가되었다. 시스템별로 전담 시스템담당자가 배정되어 있고, 전문 유지보수팀이 상주하여 업무를 진행하고 있으므로 데이터의 정합성과 직접 관련된 품질평가 지표에서는 문제가 없는 것으로 확인되었다. 그러나 유효성 규칙과 테이블 간 연결(relation)/참조 지표에서는 일부 시스템에서 부정적인 평가결과가 나왔는데, 이는 시스템 구축단계에서는 참조 관계를 정의는 했으나, 운영단계에서 절반 이상이 참조 관계를 사용하지 않는 것으로 확인되었기 때문이다. 이를 분석하기 위해 시스템 운영팀과의 인터뷰를 진행하였는데, 원인으로 확인된 것은 데이터베이스 설계 단계에서는 충실하게 테이블 간의 연계－주키와 보조키 연계, 데이터 제약조건(constraint)이 설계되었으나, 개발 단계에서의 편의성을 위하여 이러한 참조 관계를 끊었고, 이런 상황이 운영단계에까지 지속하였던 사실을 확인하였다. 이는 다른 정보시스템의 관계형 데이터베이스 운영에서도 다수 발견되는 사례인데, 시스템 테스트 단계에서 특정 케이스의 테스트시나리오를 수행하면서 전체 데이터가 변경될 경우, 다른 케이스의 시나리오 수행을 위하여 다시 데이터값을 변경 전의 값으로 돌려놓아야 하므로(roll-back), 테스트 편의를 위해 해당 케이스 외부의 참조 관계를 끊는 경우가 있기 때문이다.

KR의 데이터세트 품질평가를 통해 파악할 수 있었던 실제 데이터 품

질평가 고려사항은 다음과 같다.

첫째, 평가기관의 상황을 반영한 평가계획이 사전에 수립되어야 한다는 것이다. 이는 데이터 보유량, 데이터베이스의 구조와 특성, 품질관리 활동 수준, 품질 기준 등이 기관마다 광범하게 다르다는 상황에서 기인한 것이다. 대상 선정(전수/샘플링), 판단 기준(오류검출비율에 따른 판정) 등 구체적인 평가 원칙을 계획단계에서 충분히 검토해야 하고 그 내용에 대해서 데이터 관련 이해관계자들의 합의가 이루어져야 한다. 특히 시스템 보안 조직과의 사전협의가 반드시 필요한데, 이는 데이터세트 평가의 과정 전반에서 시스템 보안 조직의 업무협조가 없으면 데이터세트 식별부터 불가능하기 때문이다. 특히 데이터 품질평가에서는 일반 사용자의 권한보다 강한 시스템관리자 수준의 접근 권한이 필요한데, 이는 공통 데이터세트인 로그 정보 등 민감한 시스템 정보에 접근할 수 있어야 평가가 가능해진다.

둘째, 시스템 식별 단계에서 가능한 한 많은 자료를 확보해야 한다. 구축단계의 산출물과 운영 매뉴얼 등 관련 자료를 최대한 확보하고, 특히 시스템 운영 기간이 오래되었을 경우, 잦은 업데이트를 문서에 제대로 반영하지 않는 경우가 있으므로 변경되는 내용을 파악할 수 있는 자료를 확보해야 한다. 전체적인 시스템 운영에 상황을 설명해 줄 수 있는 핵심 인력을 선정하여 인터뷰 등으로 자료를 통해 확인되지 않을 수 있는 시스템의 운영맥락을 파악할 수 있어야 한다.

셋째, 실제 평가 시 해당 평가영역에 대한 의문점과 내용확인을 위해 답변하고 인터뷰할 적절한 담당자를 선정하는 것이 중요하다. 업무 가치 평가영역에 대해서는 해당 업무절차의 세부 내용과 그 절차가 전체 부서업무구성에서 어떻게 연계되어 있는지를 파악할 수 있어야 하고, 내용 가치, 연계활용 가치평가를 위해서는 데이터세트 자체가 기관 내부의 경영목표, 기관 외부의 연관기관 및 단체와 연계되어 어떤 가치를 가질 수 있는가를 파악할 수 있어야 한다. 위의 영역은 품질평가 영역

은 아니지만, 여기에서 파악된 내용 들이 다음 단계인 품질평가 단계에서 데이터 형상의 변화상을 파악하는데 귀중한 단서가 되기 때문이다. 품질평가 영역에서는 업무와 시스템담당자가 가능한 공동으로 참여해야 하고, 시스템을 실제로 관리하는 유지보수팀의 베테랑 관리자와 실무자를 참여시켜야 한다.

마지막으로 공통데이터에 해당하는 서버의 로그 정보에 대한 처분기준이 수립되어야 한다. KR 데이터세트의 웹서버, 웹 애플리케이션 서버(WAS) 로그는 행정정보시스템의 접근 내역을 확인할 수 있는 주요한 증거적 기록정보인데, 적법한 처분절차가 수립되지 않은 상태였다. 이에 따라 데이터세트 전체의 증빙적 가치의 보전을 위해서는 로그 데이터에 대한 주기적 폐기 기준과 통일된 처분지침을 수립해야 한다.

5. 결론

본 연구는 데이터세트 데이터 품질에 대하여, 기록관리 관점에서 평가할 수 있는 평가영역과 지표, 평가도구를 제시하고 실제 사례를 소개하였다. 다음은 이번 연구의 결론으로서, 데이터세트 품질평가에서 실무적으로 반드시 필요다고 생각되는 평가 원칙이다.

첫째, 이해당사자의 협업이다. 평가 대상인 데이터와 데이터세트, 행정정보시스템은 사용 주체와 운영 주체, 소유권과 이용권, 기록접근 권한이 각각 다를 수 있으며, 이에 따른 이해당사자가 다양할 수 있다. 그렇지만 데이터 품질은 개별 이해당사자의 이해관계를 초월한 기관 전체의 기록정보자산의 문제이므로 이해당사자들의 협력과 협조를 통하여 데이터 품질이라는 공동의 목표를 이루어내야 한다. 이를 위하여 기관 의사결정권자의 적극적인 지원이 필요하며, 상황에 따라 데이터 품질을 위한 태스크포스 조직을 구성할 수 있다.

둘째, 평가 대상 선정이다. 평가의 대상을 데이터세트 전체로 잡느냐(전수 평가), 혹은 표본을 선정해서 할 것이냐(샘플링 평가)의 문제이다. 상황에 따라서 데이터세트에서 중요한 메타정보—컬럼으로 구현되는—를 선택하여 데이터값이 들어있는 행 전체를 평가할 수도 있으며, 전체 컬럼을 대상으로 하되, 데이터 행(row)은 표본 조사하는 방법을 취할 수 있다.

셋째, 평가 방법이다. 평가를 자동화할 수 있는 프로파일링 도구를 개발하여 평가할 것인가, 혹은 육안 검사로 평가할 것인가 등에 대한 평가 방법을 선택해야 한다. 전자는 평가 시간을 크게 앞당길 수 있지만, 평가 규칙이 프로파일링 도구에 충실히 반영되어야 하고 도구 개발 시간이 소요되며, 데이터 구조 변경 시 도구의 업데이트까지 병행해서 진행되어야 한다. 후자는 평가가 상대적으로 자유롭게 가능하지만, 시스템전문가와 업무담당자가 협업하여야 하며, 평가를 위한 인력이 별도로 투입될 경우 반드시 충실한 교육과 전문가 검수가 필요하다.

넷째, 평가의 선순환 구조이다. 평가 결과에 따른 오류에 대해 즉각적인 대응뿐만 아니라 오류의 원인을 구조적으로 파악하여 예방과 개선방안을 마련해야 하며, 그 내용이 품질관리지침 등에 다시 반영되어야 한다. 이는 데이터세트 품질평가가 일회성에 끝나지 않고 계획－실행－평가(Plan-Do-See)의 선순환 구조를 이루어, 지속적이고 실질적인 데이터세트의 품질을 확보해야 하기 때문이다.

본 연구가 본격적으로 데이터세트 관리를 시작하고자 하는 기관에게, 부족하지만 기초가 되는 가이드가 되기를 기대한다. 그리고 기관들이 데이터세트 평가를 용이하게 수행할 수 있도록 기록관리산업계에서 공공데이터 품질평가에서 사용되는 프로파일링 툴 같은 진단 지원 도구가 개발되기를 바라며. 행정정보데이터세트 관리체계에서 등한시하던 데이터 품질 영역에 대한 관심과 향후 연구를 희망한다.

〈참고문헌〉

김유승 (2019). 디지털시대의 공공기록평가에 관한 정책적 고찰: 영국 TNA 사례를 중심으로. 한국기록학회지 62(1).5~39.

설문원 (2020). 디지털 전환시대의 공공기록정책: 기록자산으로서 정보의 관리. 한국기록학회지 63(1).5~36.

왕호성, 설문원 (2017). 행정정보데이터세트 기록의 실행방안. 한국기록관리학회지, 17(3), 23-47.

오세라, 박승훈, 임진희 (2018). 행정정보데이터세트 사례조사 연구. 한국기록관리학회지, 18(2), 109-133.

오세라, 이해영 (2019). 행정정보 데이터세트의 기록관리 방안. 한국기록관리학회지, 19(2). 51~76.

임진희, 조은희 (2010). 행정정보데이터세트 기록 이관 시 데이터 보정 및 품질개선 방법 연구. 기록학연구, (25), 91-129.

조윤희 (2004). 문화콘텐츠 통합을 위한 메타데이터 포맷 연구(Ⅱ): 도서관, 박물관, 미술관 사례를 중심으로. 한국문헌정보학회지, 38(3), 201~218.

조은희, 임진희 (2009). 행정정보데이터세트 기록의 선별기준 및 절차 연구, 기록학연구, 기록학연구, 251~291.

황진현, 박종연, 이태훈, 임진희 (2014). 행정정보시스템 기록 이관 절차와 방법 연구. 한국기록관리학회지, 14(3), 7-32.

국가기록원 (2007). 행정정보시스템 데이터세트 기록관리 방안 연구보고서.

국가기록원 (2015). 데이터세트 구조분석 및 진본성 보장 기능 모델 연구.

국가기록원 (2017). 차세대 기록관리 모델 재설계 연구개발.

국가기록원 (2020), 2020년 행정정보데이터세트 기록관리 체계 구축 최종 보고서.

국가기록원 (발간예정). 2021년 행정정보데이터세트 기록정보 서비스 및 활용모형 연구 최종보고서.

ISO 8000-2:2020(en) Data quality — Part 2: Vocabulary.

ISO/IEC 13335-1:2004 Information technology — Security techniques — Management of information and communications technology security.

ISO/IEC 38500:2015(en) Information technology — Governance of IT for the organization.

Oracle Co.(2009). SQL Developer for Database Developers An Oracle White Paper.

국가기록원 (2020). 공공기록물 관리에 관한 법률.
한국지능정보사회진흥원 (2018). 공공데이터 품질관리 매뉴얼.
한국지능정보사회진흥원 (2021). 공공데이터 예방적 품질관리 진단가이드.
행정안전부 (2019). 공공기관 데이터 자원 보존 개선방안 연구.
행정안전부 (2021). 전자정부법.
행정안전부 (2020). 공공데이터의 제공 및 이용 활성화에 관한 법률.
행정안전부 (2020). 공공데이터의 제공 및 이용 활성화에 관한 법률 시행령.
행정안전부 (2021). 공공데이터 관리지침.
행정안전부 (2021). 공공기관의 데이터베이스 표준화 지침.