OPEN ACCESS

연구데이터 품질관리를 위한 프로세스 모델 제안*

Proposal of Process Model for Research Data Quality Management

한나은 (Na-eun Han)**

초 록

본 연구는 공공데이터 품질관리 모델, 빅데이터 품질관리 모델, 그리고 연구데이터 관리를 위한 데이터 생애주기 모델을 분석하여 각 품질관리 모델에서 공통적으로 나타나는 구성 요인을 분석하였다. 품질관리 모델은 품질관리를 수행하는 객체인 대상 데이터의 특성에 따라 생애주기에 맞추어 혹은 PDCA 모델을 바탕으로 구축되고 제안되는데 공통적으로 계획, 수집 및 구축, 운영 및 활용, 보존 및 폐기의 구성요소가 포함된다. 이를 바탕으로 본 연구는 연구데이터를 대상으로 한 품질관리 프로세스 모델을 제안하였는데, 특히 연구데이터를 대상 데이터로 하여 서비스를 제공하는 연구데이터 서비스 플랫폼에서 데이터를 수집하여 서비스하는 일련의 과정에서 수행해야하는 품질관리에 대해 계획, 구축 및 운영, 활용단계로 나누어 논의하였다. 본 연구는 연구데이터 품질관리 수행 방안을 위한 지식 기반을 제공하는데 의의를 갖는다.

ABSTRACT

This study analyzed the government data quality management model, big data quality management model, and data lifecycle model for research data management, and analyzed the components common to each data quality management model. Those data quality management models are designed and proposed according to the lifecycle or based on the PDCA model according to the characteristics of target data, which is the object that performs quality management. And commonly, the components of planning, collection and construction, operation and utilization, and preservation and disposal are included. Based on this, the study proposed a process model for research data quality management, in particular, the research data quality management to be performed in a series of processes from collecting to servicing on a research data platform that provides services using research data as target data was discussed in the stages of planning, construction and operation, and utilization. This study has significance in providing knowledge based for research data quality management implementation methods.

키워드: 연구데이터 품질관리, 연구데이터 품질, 데이터 품질관리, 연구데이터 품질관리 모델, 품질관리 프로세스 모델
research data quality management, research data quality, data quality management (DQM), research data quality management model, data quality management process model

---

* 본 논문은 한국과학기술정보연구원 연구사업(과제번호: K-23-L01-C03-S01)의 지원에 의해 이루어진 것임.
** 한국과학기술정보연구원(KISTI) 박사 후 연구원(betterhan@kisti.re.kr)
■ 논문접수일자 : 2023년 2월 14일 ■ 최초심사일자 : 2023년 3월 10일 ■ 게재확정일자 : 2023년 3월 15일
■ 정보관리학회지, 40(1), 51-71, 2023. http://dx.doi.org/10.3743/KOSIM.2023.40.1.051

※ Copyright © 2023 Korean Society for Information Management
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 (https://creativecommons.org/licenses/by-nc-nd/4.0/) which permits use, distribution and reproduction in any medium, provided that the article is properly cited, the use is non-commercial and no modifications or adaptations are made.

1. 서 론

1.1 연구의 배경 및 목적

정보통신기술의 발전으로 인하여 생산되는 데이터의 양은 급속하게 증가하였고, 이미 2000년대 초반부터 데이터의 양적 보유뿐만 아니라 보유하고 있는 데이터나 정보의 질적 관리 및 활용 방안에 대한 관심 및 논의가 지속되고 있다. 데이터에 관한 질적 관리가 이루어지지 않아 부정확한 데이터를 소장 및 활용하게 되는 경우에는 데이터 품질 저하에서 유발되는 비용적 측면의 손실이 발생할 수 있고, 해당 데이터를 이용하는 이용자들의 불만족이 발생할 수 있다(정혜정, 2007; Kindling & Strecker, 2022). 이와 같은 문제 의식을 바탕으로 다양한 분야, 영역 및 기관에서 데이터 품질관리의 필요성을 인지하고 공공데이터, 빅데이터, 교육데이터 등 각 분야, 영역 및 기관의 데이터 특성에 맞는 품질관리를 실시하고 있다. 공공데이터의 경우에는 품질관리의 일환으로 품질진단을 실시하고 있으며, 진단 결과에 따르면 데이터 품질관리 수준이 높아질수록 데이터 품질오류는 낮아지는 경향이 나타난다(한국정보화진흥원, 2015).

연구데이터는 연구성과물인 연구보고서나 논문과는 구별되는 개념이며, 연구를 위해 사용된 실험재료, 과정, 결과, 관찰, 설문조사와 같은 원자료와 이를 이용하여 처리된 2차 자료를 의미하는데(한나은, 김성희, 2014), 이는 연구결과의 검증에 필수적인 것이며, 연구데이터를 공유 및 활용함으로써 연구성과를 제고하고 연구효율성 향상을 기대할 수 있게 한다(한국과학기술정보연구원, 2019). 이와 같은 중요성을 바탕으로 보조금 지원 기관 및 출판사 등은 연구데이터의 공유, 활용 및 관리를 보다 적극적으로 요청하고 있다. 국내의 경우에는 2019년, 국가연구개발사업의 관리 등에 관한 규정을 개정 및 시행하여 시범적으로 과학기술정보통신부 산하 기관에서 공고하는 과제 중 일부 과제에 대하여 데이터 관리 계획(Data Management Plan, 이하 DMP)을 적용하도록 하였으며(과학기술기본법, 법률 제18727호), 미국에서는 이미 2011년부터 미국 국립과학재단(National Science Foundation, 이하 NSF)에서 보조금 지원을 받는 연구에 한하여 DMP를 필수적으로 제출해야 함을 명시한 이후(National Science Foundation, 2014) 미국 내 대다수의 대학도서관 및 연구 기관에서 DMP를 적용하고 데이터 관리 서비스를 제공하고 있다.

이와 같이 국내・외에서 연구데이터의 공유, 활용, 관리에 대한 중요성을 인식하고 이를 활발하게 하기 위한 노력 중이며 품질관리 역시 같은 맥락에서 주목받고 있다. 그러나 연구데이터와 관련해서는 품질관리 모델보다는 관리의 측면에서 데이터 생애주기를 기반으로 한 디지털 큐레이션 혹은 데이터 큐레이션 모델 정도가 제안된 바 있다(김주섭, 신선태, 전예린, 2019). 최근 공공데이터, 빅데이터 등을 대상으로 하는 여러 품질관리 연구 및 품질관리 모델은 이미 제안된 바 있으나, 연구데이터를 대상으로 한 품질관리 프로세스 모델은 아직 제안된 바 없다. 따라서 본 연구는 연구데이터의 특성을 기반으로 한 품질관리 프로세스 모델을 제안한다는 점에서 의의를 갖는다.

1.2 연구의 내용 및 방법

본 연구는 연구데이터를 대상으로 하는 품질관리 프로세스 모델을 제안하기 위하여 연구데이터에 국한되지 않은 여러 품질관리 모델을 분석하였다. 다양한 품질관리 모델을 조사하여 비교・분석함으로써 데이터 품질관리에 필요한 공통 요인들을 추출하였고, 이를 바탕으로 연구데이터 특성에 맞는 품질관리 프로세스 모델을 제안하였다. 본 연구는 특히 연구데이터를 수집하여 서비스를 제공하는 일련의 체계에서 품질관리가 어떻게 진행될 수 있는지 뿐만 아니라 어떻게 진행되어야 하는지에 대하여 각 단계에 맞게 제안한다.

2. 이론적 배경

2.1 연구데이터의 개념 및 품질관리의 필요성

연구데이터에 대한 다양한 정의가 존재하는데, 국가연구개발혁신법 국가연구개발정보처리기준에서는 "연구데이터란 연구개발과제 수행과정에서 실시하는 각종 실험, 관찰조사 및 분석 등을 통하여 산출된 사실 자료로서 연구결과의 검증에 필수적인 데이터를 말한다"고 정의하고 있으며(국가연구개발혁신법, 법률 제18645호), 연구데이터의 형태에 관해서는 "연구데이터는 정형 및 비정형 데이터를 포함하는 형태를 가질 수 있으나 예비 분석 결과, 논문이나 저술의 초안, 연구노트, 보고서 등은 해당 되지 않는다"고 정의한다(국가과학기술연구회, 2019).

데이터의 품질 및 품질관리에 대해서는 접근방식에 따라 다양한 정의가 존재하는데, 일반적으로 데이터 품질은 "데이터를 활용하는 사용자의 다양한 활용 목적이나 만족도를 지속적으로 충족시킬 수 있는 수준"(English, 2009)이자 "비즈니스에 적합하고 정확한 데이터를 적시에 안전하고 일관성 있게 제공함으로써 비즈니스 효율을 높이고 전략적 의사결정을 지원하는 정보 자산으로서의 가치"(ISO 8000-1: 2022)로 정의한다. 이 외에도 "사용자가 사용하기에 적합하고(fit for use)"(Wang, Ziad, & Lee, 2006, 6), "사용자의 활용 목적과 요구 사항을 충족시키기 위한 데이터의 수준"(박고은, 김창재, 2015, 136)을 만족시키는 것을 데이터 품질이라고 정의한다. 이와 같은 데이터 품질을 유지, 관리 및 개선하기 위한 활동인 데이터 품질관리는 "기관이나 조직 내・외부의 정보시스템 및 DB 사용자의 기대를 만족시키기 위해 지속적으로 수행하는 데이터 관리 및 개선 활동"으로 이해할 수 있다(한국정보통신기술협회, 2022). 이를 바탕으로 본 연구에서는 품질관리의 대상인 연구데이터의 특성을 반영하여 연구데이터 품질관리를 '정형 및 비정형 데이터를 포함하는 형태의 연구데이터의 품질을 확보하고, 유지, 관리 및 개선함으로써 사용자에게 유용한 가치를 제공할 수 있도록 하는 일련의 활동'으로 정의한다.

데이터 품질관리의 의미는 데이터 생산 이후에 데이터의 결측값이나 오류를 평가하는 데이터 중심의 전통적인 품질관리 방법에서, 해당 데이터를 소유 및 운영하는 부서와 기관이 해당 데이터의 생산 단계 이전부터 데이터의 특성에 기반한 향후의 활용 방안까지를 고려하여

수행하는 방향으로 확장되고 있다(송치호, 임진희, 2021). 전통적인 접근법인 데이터 중심의 품질관리는 데이터의 오류를 측정하고 발견된 오류를 수정하는 방식으로 진행되는데, 이는 데이터 오류를 신속히 개선하고 데이터 품질을 계량적으로 제시할 수 있다는 점에서 강점을 가지나, 향후에 같은 종류의 오류가 반복적으로 발생할 수 있기 때문에 근본적인 품질 개선 방법이라고 보기 어렵다(김선호, 이창수, 2013). 반면에 프로세스 중심의 품질관리는 데이터 품질관리의 프로세스를 유지 및 개선하는 방법으로써 데이터의 오류뿐만 아니라 프로세스 자체를 개선하는 방식이다. 프로세스 중심 품질관리를 적용함으로써 데이터 오류의 근본적인 원인을 제거하고, 동일한 데이터 오류의 재발 방지를 기대할 수 있다(김선호, 이창수, 2013).

연구데이터 공유의 중요성이 증대하면서 연구데이터 품질관리의 필요성 역시 대두되고 있다. 연구데이터는 연구개발과제 수행 과정에서 산출된 사실 자료로서 연구결과에 대한 핵심 정보를 포함하고 있으며, 연구결과 검증에 필수적인 데이터이다(국가연구개발혁신법, 법률 제18645호). 전세계적으로 데이터 집중형 과학, 데이터 집중형 연구로 흐름이 변화함에 따라 연구데이터의 공유, 관리, 재활용에 대한 관심이 높아지고 있으며, 연구데이터를 공유함으로써 연구자들의 중복 연구 방지 및 데이터 중복 생산 방지, 연구결과 검증, 빠른 지식 공유 등을 기대할 수 있게 되었다(한국과학기술정보연구원, 2019). 데이터 공유의 중요성과 더불어 데이터 자체의 품질관리의 중요성에 대한 인식도 높아지고 있는데, 데이터의 양적 증가 및 공유에 대한 관심이 높아짐에 따라 고품질 데이터를 제공받고자 하는 이용자들의 요구 및 기대도 증가하고 있다. 연구데이터의 품질관리를 실시함으로써 데이터 품질 진단을 통한 소장 및 연계데이터의 정확한 수준을 파악할 수 있고, 데이터 품질 진단을 기반으로 하는 데이터 오류의 근본적 해결 방안을 도출할 것을 기대할 수 있다. 뿐만 아니라 일회성 시스템 개선으로는 해결이 어려운 근본적 데이터 오류 문제 해결을 기대해볼 수 있으며, 데이터 오류 감소 및 데이터의 신뢰성 증대를 바탕으로 하는 고품질의 연구데이터 유지가 가능하다. 제공되는 고품질의 데이터는 이용자들의 만족도 향상 및 분석 시스템의 신뢰도 향상, 그리고 각각의 데이터를 통합하는 경우에 시간 절약이 가능하다는 측면 등 다양한 이점들을 제공할 수 있다(Eckerson, 2002).

2.2 데이터 품질관리 지표

데이터의 품질을 관리 및 측정하기 위해 다양한 관리지표 및 평가지표 등이 제안된 바 있다. 데이터 품질지표는 다수의 정보 수요자가 지속적으로 요구하는 데이터 품질속성들의 집합이며, 이 데이터 품질지표의 모음에 따라 데이터 품질이 정의될 수 있다(Wang, Ziad, & Lee, 2006). 데이터 품질속성은 데이터 품질특성, 품질기준, 품질요소 등과 같은 용어와 혼용되어 사용되어지는 경우가 많으나, 본 연구에서는 데이터 품질관리를 목적으로 데이터 품질의 관리 및 측정에 사용됨을 감안하여 품질기준이라는 용어를 사용하도록 한다. 데이터 품질기준은 품질을 관리 및 측정하고자 하는 데이터의 특성에 따라 다양하게 정의되어 활용되

고 있다.

한국데이터베이스진흥원의 DQC(Database Quality Certification)는 데이터 특성에 따라 정형데이터와 비정형데이터를 구분하여 품질기준을 정의하여 제시하고 있다(한국데이터베이스진흥원, 2006). 정형데이터는 구조화된(structured) 데이터로, 데이터베이스에 저장된 구조적 데이터를 의미하며 대표적인 예로 스프레드시트를 들 수 있다. 반면에 비정형데이터는 비구조화된(unstructured) 데이터로써 형태가 없고, 연산이 불가능한 데이터이다. 비정형데이터는 동영상, 이미지, 음성, 텍스트 등을 포함한다. DQC 품질기준에 따르면 정형데이터는 완전성, 유일성, 유효성, 일관성, 정확성을 만족하여야 하고, 비정형데이터는 기능성, 신뢰성, 사용성, 효율성, 이식성을 만족할 필요가 있다(한국데이터베이스진흥원, 2006). 또 다른 품질기준인 안전행정부 품질관리 매뉴얼의 품질기준에 따르면 정형데이터, 비정형데이터에 대한 구분없이 준비성, 완전성, 일관성, 정확성, 유효성, 보안성, 적시성, 유용성의 총 8가지 기준을 충족시켜야 한다(안전행정부, 2014). 가장 최근에 제안된 품질기준은 2022년 CODATA(Committee on Data for Science and Technology)의 연구 논문에서 제시되었으며, 기존의 연구에서 제안된 내용들을 분석 및 조합하여 데이터 품질기준을 제시하였다(Kindling & Strecker, 2022). 해당 연구에서는 총 10가지의 품질기준이 제시되었는데, 이는 접근성, 정확성, 사용 방법의 적절성, 메타데이터의 적절성, 완전성, 일관성, 포괄성, 데이터 포맷의 개방성, 데이터 라이센스의 개방성, 재사용 가능성을 포함한다(Kindling & Strecker, 2022).

해당 내용을 표로 정리하면 다음 〈표 1〉과 같다.

이외에도 김형섭(2020)은 데이터 품질관리의 중요도를 논의하면서 품질기준에서 정확성, 일관성, 보안성, 완전성, 준비성 순으로 중요도가 있음을 주장한 바 있다.

3. 데이터 품질관리 모델 분석

본 연구는 데이터 품질관리를 위해 제안된 모델 및 연구데이터의 관리를 목적으로 하여 제안된 데이터 생애주기 모델을 조사 및 분석하였다. 데이터 품질관리 모델은 최근 5년 이내에 특정 데이터를 대상으로 품질관리를 실시하기 위해 모델을 제안하고 매뉴얼 형태로 품질관리 지침을 제시한 두 개의 품질관리 모델을 대상으로 선정하였고, 추가적으로 품질관리에 초점이 맞추어진 것은 아니지만 품질관리를 포함하는 관리를 목적으로 제안된 연구데이터 생애주기 모델을 대상으로 선정하여 분석하였다.

본 연구는 총 3개의 데이터 품질관리 모델을 분석하였으며, 이는 2018년 한국정보화진흥원에서 제안한 공공데이터 품질관리 모델, 2021년 한국정보사회진흥원에서 제안한 빅데이터 품질관리 모델, 그리고 2019년 김주섭, 김선태, 전예린이 제안한 연구데이터 관리를 위한 데이터 생애주기 모델을 포함한다.

3.1 공공데이터 품질관리 모델

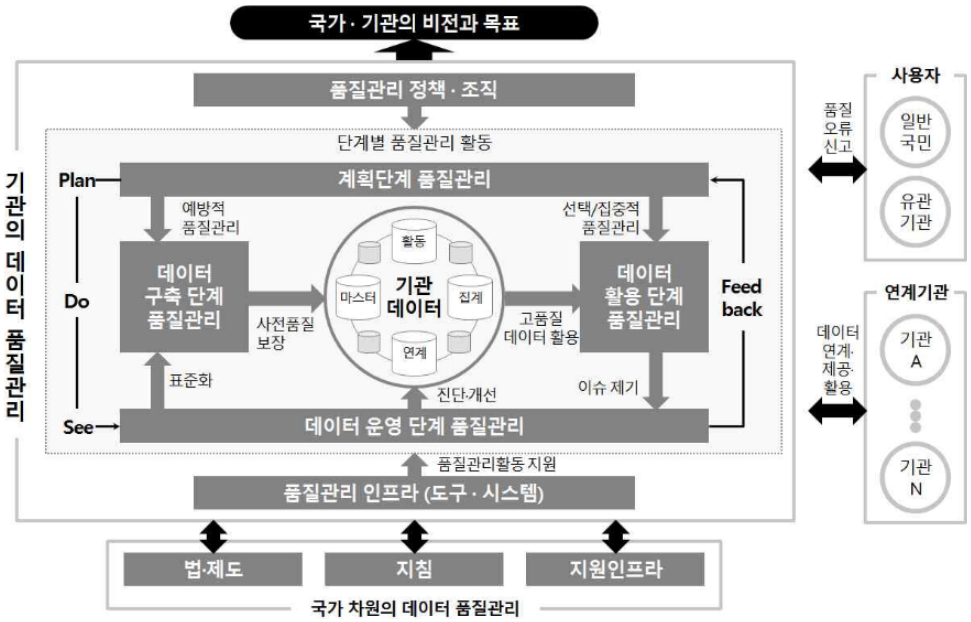
공공데이터 품질관리 모델은 공공기관이 수행해야 할 데이터 품질관리 전반에 대한 구성

〈표 1〉 데이터 품질기준 비교

	DQM	안전행정부	CODATA	정의
품질지표	완전성	완전성	완전성	• 필수항목에 누락이 없고 필요한 모든 구성 요소의 데이터가 존재함
	일관성	일관성	일관성	• 데이터의 속성이 일관되며 형태가 서로 일치함
	정확성	정확성	정확성	• 데이터는 설명하고자 하는 바를 분명하고 정확하게 나타냄
	유효성	유효성		• 데이터 항목은 정해진 데이터의 유효 범위 및 도메인을 충족해야 함
	사용성		사용방법의 적절성	• 사용 방법이 데이터의 수집, 처리 및 제공에 적절하고 올바르게 적용됨
	기능성			• 사용 목적에 부합해야 하며, 요구 기능을 충족시켜야 함
	신뢰성			• 데이터 사용시 정해진 신뢰수준을 유지하고, 사용자의 오류를 방지해야 함
	효율성			• 사용되는 자원에 비해 발휘되는 성능을 제공하는 정도가 일정 수준에 도달해야 함
	이식성			• 해당 데이터가 다양한 환경 및 상황에서 실행될 수 있어야 함
	유일성			• 데이터 항목은 유일해야 하며 중복되어서는 안됨
		보안성		• 데이터 생성관리 주체 관리수준, 권한에 따른 데이터 접근관리수준, DB보안관리 수준 등을 유지해야 함
		준비성		• 데이터 품질관리 요소에 대한 정의, 지속적 활동 및 최신성을 유지해야 함
		적시성		• 사용자 만족 수준에 달하는 응답시간 및 정보요구사항을 만족하고, 작업시간의 최소화 및 정보의 최신성을 유지해야 함
		유용성		• 사용자가 만족하는 수준의 충분한 정보를 제공하고, 정보 접근의 사용자 편의성을 만족시켜야 함
			접근성	• 최소한의 단계로 데이터에 접근할 수 있도록 해야 함
			메타데이터의 적절성	• 메타데이터는 데이터를 적절하게 설명할 수 있어야 함
			포괄성	• 데이터는 얼마간의 시간적 공간적 범위를 포괄하는 범위를 가져야 함
			데이터 포맷의 개방성	• 데이터는 공개적이고 비독점적인 형식으로 제공되어야 함
			데이터 라이센스의 개방성	• 데이터는 오픈 라이선스로 제공되어져야 함
			재사용 가능성	• 데이터는 향후 다른 사람을 통해 분석 및 활용될 수 있어야 함

체계를 제시하고, 국가 차원의 품질정책 및 데이터 생애주기를 고려하여 기관의 데이터 품질관리를 위한 계획, 구축, 운영, 활용의 각 단계에서 품질관리 활동을 구분하고 해당 단계에서 중점적으로 고려해야할 데이터 품질관리 사항에 대해 설명한다. 제안된 공공데이터 품질관리 모델은 다음 〈그림 1〉과 같으며, 해당 모델은 기관의 데이터 품질관리 정책과 이를 수행하기 위한 조직 체계, 기관이 보유하고 있는 데이터의 품질수준 제고를 위한 계획, 구축, 운영, 활용 단계별 품질관리 활동을 지원하기 위한 품질관리 인프라로 구성된다(한국정보화진흥원, 2018).

공공데이터 품질관리 모델은 계획단계, 구축단계, 운영단계, 그리고 활용단계인 총 4단계로 구성되어 있는데, 이는 정보생명주기(Information Life Cycle)를 바탕으로 품질관리 체계를 고안



〈그림 1〉 공공데이터 품질관리 모델(한국정보화진흥원, 2018, 25)

한 것이다(한국정보화진흥원, 2018). 정보생명주기는 데이터가 처음 생성되어 보존 및 폐기되는 일련의 단계라고 이해할 수 있는데, 일반적으로 계획(Plan), 획득(Obtain), 저장 및 공유(Store & Share), 운영(Maintain), 활용(Apply), 그리고 폐기(Dispose)의 단계로 구성되어 있다.

공공데이터 품질관리 모델을 각 단계별로 살펴보면, 계획단계는 기관이 정보화 계획을 수립하는 단계에서 고려해야 하는 데이터 품질관리 측면의 계획 수립 활동을 실시하는 단계이다. 해당 단계에서는 기관 데이터베이스 품질관리 조직 및 인력, 기관 데이터베이스 품질 목표 정의, 기관 중점 데이터베이스 품질관리 대상 선정, 기관 데이터베이스 품질 진단 및 개선 계획, 기관 데이터베이스 표준화 방안, 연계데이터 품질확보 방안뿐만 아니라 그 밖의 기관 데이터베이스 품질관리를 위해 필요한 사항들을 고려해야 한다(한국정보화진흥원, 2018).

구축단계는 기관의 데이터베이스 및 정보시스템의 신규 도입 또는 고도화 등을 추진하기 위한 사업의 구축단계에서 고려해야 할 데이터 품질관리 활동을 의미하는데, 해당 단계에서는 데이터 표준화와 데이터 산출물 관리를 집중적으로 고려해야 한다(한국정보화진흥원, 2018). 데이터 표준화는 기관이 보유 및 운영하고 있는 시스템에 산재되어 있는 데이터 정보 요소에 대해서 명칭, 정의, 형식, 규칙 등에 대한 원칙을 수립하고, 이를 기관 전체의 데이터로 적용하는 것인데, 이와 같은 데이터 표준화의 목적은 기관 차원에서 사용하는 용어의 의미와 형식에 규칙을 정함으로써 의사소통의 혼란을 방지하고 데이터의 정확성 및 일관성을 유지하여 궁극적으로 고품질의 공공데이터를 확보하는 것에 있다. 다음으로 데이터 산출물 관리는 정보시스템

의 신규 도입과 시스템 고도화 등의 구축단계에서 생성되는 산출물들 가운데 데이터 품질과 관련된 산출물들을 생성 및 검증하고 현재까지의 진행 사항을 반영하는 일련의 활동으로 이해할 수 있다(한국정보화진흥원, 2018).

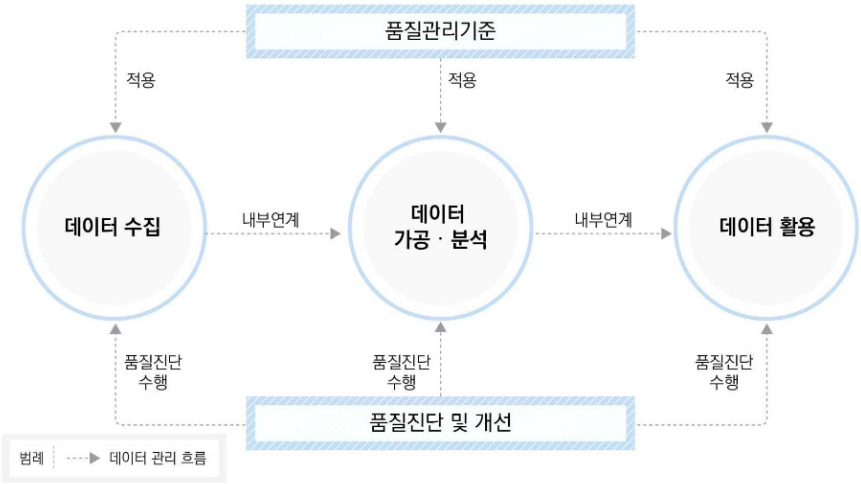
다음으로 운영단계의 품질관리는 기관이 생성, 보유 및 활용하고 있는 데이터를 운영하는 단계에서 데이터의 품질 수준을 향상시키기 위한 제반 활동이라고 이해할 수 있다. 해당 단계에서는 연계데이터에 대한 품질 관리, 품질관리 계획 수립 단계에서 선정된 데이터베이스에 대한 품질 진단 및 개선, 그리고 데이터베이스 산출물 점검, 데이터 변경에 따른 문서 최신성 확보 및 이해관계자에게 변경사항을 통지하는 활동들이 진행되어야 한다(한국정보화진흥원, 2018).

마지막으로 활용단계 품질관리는 기관 내·외부의 사용자가 데이터를 활용하는 중 발생할 수 있는 데이터 품질오류 신고를 관리하고 데이터 활용성과를 평가하는 일련의 활동이라고 이해할 수 있다. 해당 단계는 데이터를 활용함에 있어서 발생할 수 있는 품질 이슈를 인지하여 개선할 수 있도록 연계하는 품질오류 신고관리 활동과 품질 개선 결과를 포함하는 데이터 품질진단 및 평가에 따른 조치의 활동으로 구성된다(한국정보화진흥원, 2018). 먼저 품질오류 신고관리는 일반 이용자와 유관기관을 포함하는 데이터 사용자로부터 품질 오류에 대한 신고를 접수 받고 이를 확인 및 조치하여 개선결과를 통지하는 과정으로 구성되며, 이 활동의 목적은 기관 내부에서 자체적으로 인지하지 못한 데이터의 오류를 인식하고 이를 개선하여 제공함으로써 기관이 보유한 데이터의 품질 수준을 향상시키는 것에 있다. 다음으로 데이터 품질진단 및 평가에 따른 조치는 데이터의 품질 수준을 진단 및 평가하고 그 결과에 따라서 시정조치 계획을 수립하고 문제를 개선함으로써 데이터 이용자의 요구에 맞는 데이터 품질을 확보하기 위한 일련의 활동으로 이해할 수 있다(한국정보화진흥원, 2018).

3.2 빅데이터 품질관리모델

빅데이터 품질관리모델은 빅데이터 플랫폼 및 센터에서 데이터 품질관리 활동을 수행하는데 활용할 수 있도록 제안되었으며, 데이터 수집, 분석, 유통 체계에서 데이터베이스의 품질 향상을 위해 어떤 점을 고민해야 하고, 어떤 방식으로 수행해야 하는지 이해를 돕기 위해 고안되었다(한국지능정보사회진흥원, 2021). 구체적인 가이드 없이 데이터 품질관리를 실시하는 경우에는 전반적인 데이터 품질관리에 대한 이해의 부족으로 인해서 시행착오 및 양질의 데이터 확보 실패 등의 문제가 발생할 수 있다. 이와 같은 이해를 바탕으로 빅데이터 품질관리모델은 데이터 품질관리에 대한 이해를 고취시키고 구체적인 품질관리 기법을 제공함으로써 빅데이터 플랫폼 및 센터에서 보다 효율적인 데이터 품질관리 활동을 수행할 수 있도록 하는 것을 목적으로 한다(한국지능정보사회진흥원, 2021).

빅데이터 품질관리모델은 빅데이터 생애주기를 고려하여 데이터 수집, 데이터 가공·분석, 데이터 활용의 관리 흐름을 바탕으로 제안되었을 뿐만 아니라 PDCA(Plan-Do-Check-Act) 모델의 흐름을 포함하고 있다. 빅데이터 플랫폼 및 센터에서 제안한 빅데이터 품질관리모델은 다음 〈그림 2〉와 같다.



〈그림 2〉 빅데이터 품질관리모델(한국지능정보사회진흥원, 2021, 44)

빅데이터 품질관리모델은 빅데이터 생애주기에 맞추어 데이터 수집, 데이터 가공・분석, 데이터 활용으로 구성되어 있는데, 각 단계에서는 적용, 품질 진단 수행 및 내부 연계를 실시한다(한국지능정보사회진흥원, 2021). 가장 먼저, 데이터 수집 단계에서는 품질 진단 수행을 위한 데이터 품질관리기준을 적용하게 되며, 품질진단 수행을 통해 수집된 데이터에 대하여 적용된 데이터 품질관리기준을 바탕으로 품질진단 활동을 진행한다. 내부연계는 데이터 가공 및 분석을 위해 품질진단 및 개선이 완료된 수집 데이터에 대하여 내부 연계 송신을 진행하는 것을 의미한다(한국지능정보사회진흥원, 2021).

데이터 가공・분석 단계에서의 적용은, 데이터 가공 및 분석 단계에 품질진단 수행을 위한 데이터 품질관리기준을 적용하는 것을 의미한다. 이후 가공된 데이터에 대하여 적용된 데이터 품질관리기준을 바탕으로 품질진단을 수행하게 되며, 해당 단계에서의 내부연계는 데이터 가공 및 분석을 위해 품질진단 및 개선이 완료된 수집 데이터에 대한 내부 연계 수신과 데이터 활용을 위해 품질 진단 및 개선이 완료된 가공 데이터에 대한 내부 연계 송신을 의미한다(한국지능정보사회진흥원, 2021).

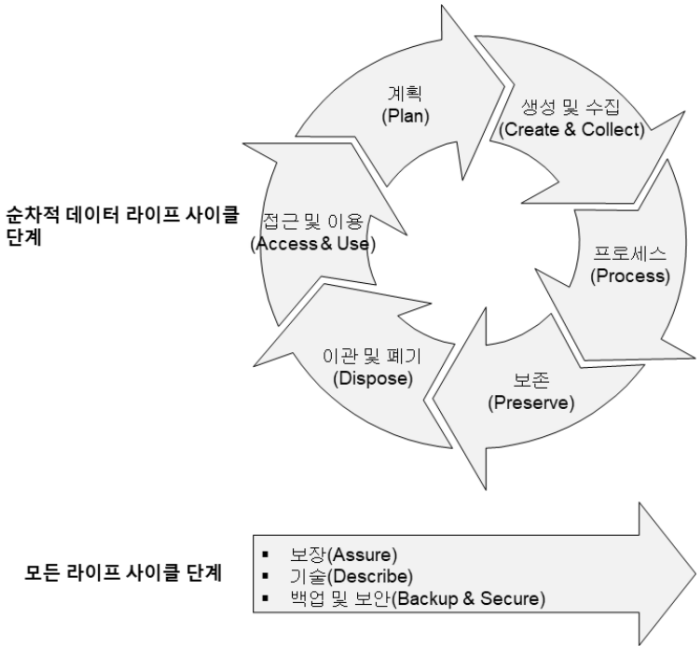
마지막으로 데이터 활용 단계에서의 적용은 데이터 활용 단계에 품질 진단 수행을 위한 데이터 품질관리기준을 적용하는 것을 의미하며, 품질진단 수행은 활용을 위해 저장된 데이터에 대하여 적용된 데이터 품질관리기준을 바탕으로 품질진단을 실시하는 것이다. 해당 단계에서의 내부연계는 데이터 활용을 위해 품질진단 및 개선이 완료된 가공 데이터에 대해 내부 연계로 데이터를 수신하는 것을 의미한다(한국지능정보사회진흥원, 2021).

3.3 연구데이터 생애주기 모델

연구데이터 생애주기 모델은 기존에 제안된 데이터 생애주기 모델을 분석하여 구성 요소를 추출하고 해당 요소의 세부 내용을 중심으로 문

헌 분석을 실시하여 연구데이터 관리 및 서비스 지원을 목적으로 제안되었다. 제안된 생애주기 모델은 2개의 영역으로 구분되는데 첫째는 순차적 데이터 생애주기 단계를 나타낸 것으로 계획(Plan), 생성 및 수집(Create & Collect), 프로세스(Process), 보존(Preserve), 이관 및 폐기(Dispose), 접근 및 이용(Access & Use) 단계가 포함되며, 둘째는 모든 생애주기 단계를 나타낸 것으로 보장(Assure), 기술(Describe), 백업 및 보안(Backup & Secure) 단계가 포함된다(김주섭, 김선태, 전예린, 2019). 해당 연구데이터 생애주기 모델은 다음 〈그림 3〉과 같다.

연구데이터 관리를 위한 생애주기 모델의 단계별 세부 내용을 살펴보면, 순차적 단계인 '계획'단계에서는 연구의 제안서 및 DMP를 작성하고, 데이터 및 메타데이터의 형식 및 품질에 대한 표준을 식별한다. '생성'단계에서는 데이터를 생성하기 위한 가장 좋은 방법을 결정 후 데이터를 생성하며 동시에 메타데이터를 할당하고 생성하며, '수집'단계에서는 기존의 데이터를 탐색 및 확보하는데, 정량적 데이터와 정성적 데이터를 수집하며 데이터의 작성자, 아카이브, 리포지토리 또는 데이터 센터를 통해 데이터를 수집하게 된다. '처리'단계에서는 데이터의 입력, 검증, 변환, 통합, 설명, 요약 등을 통해 연구데이터를 처리하고, '분석'단계에서는 통계 분석을 포함하여 시각화, 이미지 분석 및 공간 분석, 모델링을 실시하고 데이터 분석을 위한 소프트웨어를 제시한다. '보존'단계에서는 데이터의 관리 특성을 장기간 보존하고 보유하기 위한 조치를 실시하는데, 데이터를 장기적으로 보존할 계획을 수립하며 보존을 위한 메



〈그림 3〉 연구데이터 관리를 위한 생애주기 모델(김주섭, 김선태, 전예린, 2019, 335)

타데이터 및 자료를 생성한다. '이관 및 폐기'단계에서는 데이터를 폐기할 것인지 혹은 다른 조직으로 전송할 것인지 결정하게 되는데, 큐레이션 및 보존을 위해 선택되지 않은 데이터는 폐기시키도록 하며, 혹은 다른 아카이브, 리포지토리, 데이터 센터 등으로 이관한다. '접근 및 이용'단계에서는 지정된 이용자와 재사용자 모두가 항상 데이터에 접근이 가능한지를 확인해야 하는데, 접근 제어 및 인증 절차가 적용되며 이때 사용되는 파일 형식을 제시해야 한다. 순차적 단계의 마지막인 '출판'단계에서는 데이터에 대한 정보를 배포하게 되는데 해당 단계에서 데이터를 인용, 배포, 공유하고 저작권을 설정하게 된다(김주섭, 김선태, 전예린, 2019).

모든 생애주기 단계는 기술, 보장, 백업 및 보안 단계로 구성되는데, '기술'단계에서는 메타데이터를 할당하고 생성 및 유지하도록 할 뿐만 아니라 데이터의 사용자, 시기, 위치, 이력, 활용 방법 등을 명기하여 데이터를 문서화하도록 한다. 다음으로 '보장'단계에서는 데이터의 무결성을 유지하면서 진본, 신뢰성 및 가용성을 보장하도록 하는데, 유효성을 검사하고, 품질 보증 및 품질 관리를 실시하며 데이터 품질을 문서화한다. '백업'단계에서는 데이터를 관련 표준을 준수하는 안전한 방식을 활용하여 저장하며, 잠재적인 손실을 최소화하기 위해서 단기적 보존 계획을 수립하고 적용한다. 마지막으로 '보호'단계에서는 접근 제한을 통한 개인 정보 보호를 실시하고 기밀성을 보호하도록 하며 우발적인 데이터 손실 및 손상, 그리고 무단 접근으로부터 데이터를 보호할 수 있도록 한다(김주섭, 김선태, 전예린, 2019).

3.4 데이터 품질관리모델 비교분석

본 연구는 데이터 품질관리를 위해 제안된 공공데이터 품질관리 모델과 빅데이터 품질관리 모델, 그리고 연구데이터의 관리를 목적으로 하여 제안된 데이터 생애주기 모델을 조사 및 분석하였다. 본 연구에서 분석한 각각의 모델은 특정 데이터의 품질관리를 목적으로 하고 있으며 생애주기를 기반으로 혹은 데이터 품질관리 체계에 맞추어 제안되었다. 각 모델의 품질관리 대상, 체계, 구성요소를 정리하면 다음 〈표 2〉와 같다.

각 품질관리 모델은 품질관리의 목적에 따라 공공데이터, 빅데이터, 연구데이터를 대상으로 하는 품질관리 혹은 생애주기 기반의 데이터 관

〈표 2〉 데이터 품질관리모델 비교

	품질관리 대상 데이터	품질관리 체계	품질관리 구성요소
공공데이터 품질관리 모델	공공데이터	생애주기/ 품질관리 프로세스	계획, 획득, 저장 및 공유, 운영, 활용, 폐기
			계획, 구축, 운영, 활용
빅데이터 품질관리 모델	빅데이터	PDCA/ 품질관리 프로세스	계획, 실천, 수행, 보완
			데이터 수집, 가공 및 분석, 활용
연구데이터 생애주기 모델	연구데이터	생애주기	계획, 생성 및 수집, 프로세스, 보존, 이관 및 폐기, 접근 및 이용, 보장, 기술, 백업 및 보안

리를 목적으로 하여 제안되었고, 품질관리 체계는 데이터의 생애주기, 품질관리 프로세스, 혹은 PDCA 모델을 기반으로 하여 만들어졌음을 알 수 있다. 각 모델의 품질관리 구성요소를 살펴보면 공공데이터 품질관리 모델은 데이터 생애주기 및 품질관리 프로세스를 바탕으로 제안되었는데, 데이터 생애주기는 계획, 획득 저장 및 공유, 운영, 활용, 폐기의 요소들로 구성되어 있고, 품질관리 프로세스는 계획, 구축, 운영, 활용 총 4단계로 진행된다. 빅데이터 품질관리 모델은 PDCA 모델과 품질관리 프로세스를 바탕으로 제안되었으며 품질관리 구성요소로 계획(Plan), 실천(Do), 보완(Check), 수행(Act)과 데이터 수집, 데이터 가공 및 분석, 데이터 활용의 체계를 포함한다. 연구데이터 생애주기 모델은 데이터 생애주기를 기반으로 관리를 위한 생애주기 모델을 제안하였는데, 해당 모델의 구성요소는 계획, 생성 및 수집, 프로세스, 보존, 이관 및 폐기, 접근 및 이용, 보장, 기술, 백업 및 보안을 포함한다.

각각의 품질관리 모델의 구성요소들은 데이터의 특성 및 관리 체계에 따라 차별적인 요인들이 존재하지만, 동시에 데이터 품질관리 프로세스에서 공통적으로 자리하는 요인들이 존재한다. 본 연구에서는 계획, 수집 및 구축, 운영 및 활용, 보존 및 폐기의 구성요소를 데이터 품질관리 모델에서 공통적으로 활용하는 요인으로 추출하였다.

4. 연구데이터 품질관리 프로세스 모델 제안

앞의 3장에서 여러 종류의 데이터 품질관리 모델들을 분석함으로써 품질관리 체계 및 품질관리 구성요소들을 살펴보았으며, 공통적으로 품질관리 모델에서 계획, 수집 및 구축, 운영 및 활용, 보존 및 폐기의 구성요소를 활용하고 있음을 분석하였다. 본 장에서는 앞서 분석한 내용을 바탕으로 연구데이터를 대상으로 하여 계획, 구축 및 운영, 활용 단계로 구성된 품질관리 프로세스 모델을 제안하며, 이를 앞 장에서 정리하여 제시한 〈표 2〉와 같이 모델의 체계 및 구성요소로 정리하면 다음 〈표 3〉과 같다.

이미 언급한 바와 같이 프로세스 중심의 품질관리는 전통적 접근법인 데이터 중심의 품질관리와는 다른 개념이며, 데이터 품질관리의 프로세스를 유지 및 개선함으로써 데이터의 오류 뿐만 아니라 프로세스 자체를 개선하도록 하는 방식이다(김선호, 이창수, 2013). 본 연구에서는 공공데이터 품질관리 모델의 체계를 참고하되, 특히 연구데이터를 대상 데이터로 하여 서비스를 제공하는 연구데이터 리포지토리 및 연구데이터 서비스 플랫폼에서 데이터를 수

〈표 3〉 연구데이터 품질관리 프로세스 모델 체계 및 구성요소

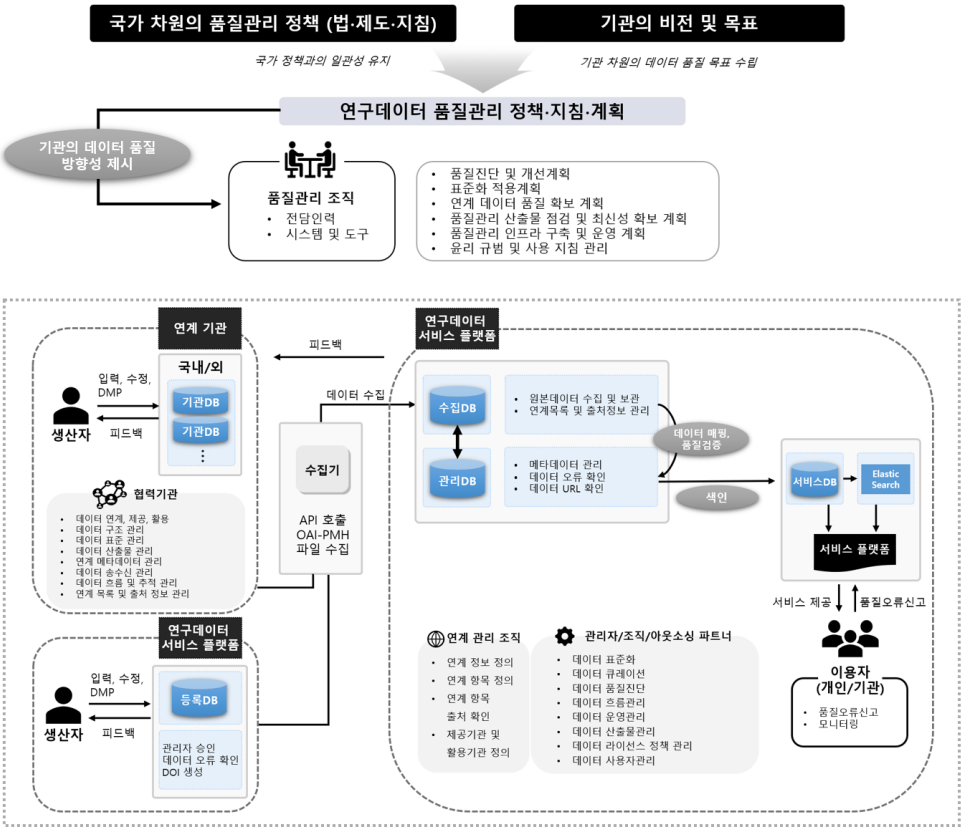
	품질관리 대상 데이터	품질관리 체계	품질관리 구성요소
연구데이터 품질관리 프로세스 모델	연구데이터	생애주기/ 품질관리 프로세스	계획, 생성 및 수집, 운영, 저장 및 공유, 보존, 폐기
			계획, 구축 및 운영, 활용

집하여 서비스하는 일련의 과정에서 어떠한 품질관리가 이루어질 수 있으며 이루어져야 하는지에 대해 논의한다.

연구데이터 품질관리를 수행하기 위해서는 기관 차원에서의 데이터 품질 목표를 달성하기 위한 정책, 지침, 계획을 수립하여야 하고, 이를 수행하기 위한 조직체계를 정의 및 구축하고 관리할 필요가 있다. 본 연구에서 제안하는 연구데이터 품질관리 프로세스 모델은 계획활동, 구축 및 운영활동, 활용단계로 나누어 각 목표에 맞는 활동을 제시한다. 각 단계별 품질관리를 포함하여 전반적인 연구데이터 품질관리 프로세스를 도식화하여 모델로 제시하면 다음 〈그림 4〉와 같다.

4.1 계획단계 품질관리

계획단계 품질관리는 연구데이터 품질관리 계획을 수립하는 단계에서 고려해야 할 행동을 의미하며 연구데이터 품질관리는 국가 차원의 품질관리 정책과 맥락을 같이 하고, 기관 차원의 데이터 품질 목표를 달성하기 위한 정책, 지침, 계획을 수립하는 것을 포함한다. 연구데이터 품질관리 계획을 수립하는 단계에서 고려해



〈그림 4〉 연구데이터 품질관리 프로세스 모델

야 할 주요 내용은 품질진단 및 개선계획, 표준화 적용계획, 연계데이터 품질 확보 계획, 품질관리 산출물 점검 및 최신성 확보 계획, 품질관리 인프라 구축 및 운영 계획, 윤리 규범 및 사용 지침 관리와 같다(〈그림 4〉 참조).

품질진단 및 개선계획은 품질관리 대상 데이터베이스 및 데이터를 선정하고, 선정된 데이터베이스 및 데이터에 대한 품질진단 및 개선계획을 수립하는 것이라고 할 수 있는데, 장·단기적 계획을 바탕으로 당해 계획을 수립하는 것이 필요하다. 예를 들어 품질진단 계획 혹은 품질개선 계획만을 수립할 수 있고, 품질진단 계획과 품질개선 계획을 모두 수립할 수도 있다. 표준화 적용계획은 표준화 목표에 따라 표준화 대상 데이터베이스 및 데이터의 유형을 선정하고 표준화 적용 범위 등에 대한 계획을 수립하는 것이며, 연계데이터 품질 확보 계획은 외부 기관으로부터 수집한 데이터를 이용자들에게 제공하는데 있어서 발생할 수 있는 문제점을 검토하고 해당 기관과 지속적인 데이터 품질 이슈에 대한 논의와 개선 방안을 계획하는 것이다. 이는 현재 연계 중인 기관 뿐만 아니라 신규 연계 기관을 포함하는 계획 수립 방안이 필요한 부분이다. 다음으로 품질관리 산출물 점검 및 최신성 확보 계획은 품질관리 대상 산출물이 규정 및 지침대로 관리되어지고 있는지 뿐만 아니라 산출물의 최신성 확보가 원활하게 이루어지고 있는지를 점검하기 위한 계획을 수립하는 것을 의미한다. 품질관리 인프라 구축 및 운영 계획은 데이터 품질관리 활동을 효과적으로 지원할 수 있는 품질관리 인프라의 도입 및 운영에 관한 계획을 수립하는 것인데, 해당 리포지토리 규모에 대한 이해와 품질관리 시스템 도입의 필요성 등에 대한 논의가 선행되어야 한다. 계획단계의 마지막은 윤리 규범 및 사용 지침 관리인데 이는 개인정보보호법 등의 정보보호 관련 법규를 반영해야 할 뿐만 아니라 기관의 품질관리 정책을 주체화한 지침을 수립하고 관리해야 한다는 것을 의미한다.

4.2 구축 및 운영단계 품질관리

구축단계 품질관리는 기관의 정보시스템의 신규 도입 혹은 고도화 등을 추진하기 위한 구축사업 단계에서 고려해야 할 데이터 품질관리 활동을 의미하는데(한국정보화진흥원, 2018), 운영단계 품질관리는 기관이 생성, 보유, 활용하고 있는 데이터를 운영하는데 있어서 데이터의 품질 수준을 향상시키기 위한 제반 활동을 의미한다. 구축 및 운영단계 품질관리에서 고려해야 하는 주요 활동은 데이터 표준화, 데이터 산출물 관리, 연계데이터에 대한 품질관리, 품질관리 계획 수립단계에서 선정된 데이터베이스 및 데이터에 대한 품질진단 및 개선활동이다(〈그림 4〉 참조).

연구데이터는 연구데이터를 서비스하는 플랫폼을 통해 직접 입력되거나 외부 기관과의 연계를 통해 수집되며, 모든 입력데이터는 수집기를 통해 수집DB로 넘어가게 된다. 수집DB에서는 원본데이터를 수집 및 보관하며 연계목록 및 출처정보를 관리할 필요가 있으며 이를 관리DB로 이관하는 과정에서 데이터 매핑과 품질검증을 수행해야 한다. 관리DB에서는 데이터 오류 및 URL 등이 제대로 작동하는지를 확인하고 메타데이터를 검증 및 관리하여야 하며, 해당 데이터가 서비스DB로 넘어가는 과정에서 색인이 진행

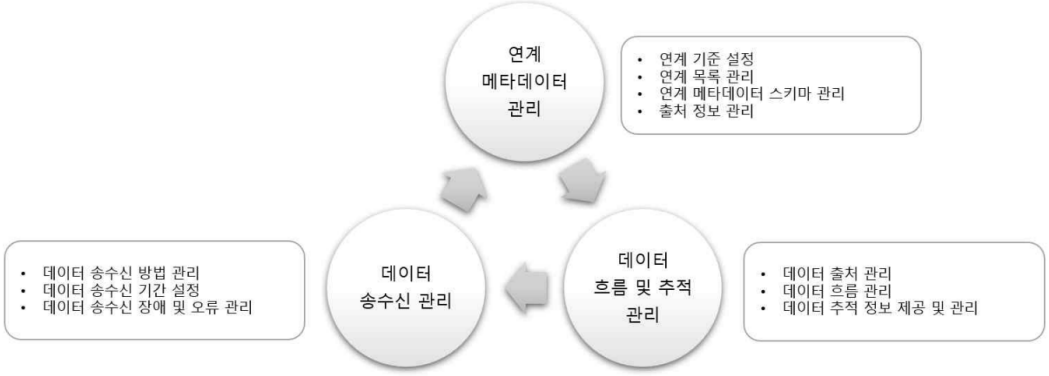
될 수 있도록 한다. 이렇게 정제된 연구데이터는 서비스DB와/또는 ElasticSearch를 거쳐 연구데이터 서비스 플랫폼에서 제공될 수 있다.

연계데이터 품질 확보를 위해서는 연계기관과의 긴밀한 협조가 요구되며, 연계데이터와 관련하여 메타데이터 관리, 데이터 흐름 및 추적 관리, 데이터 송수신 관리 등이 요구된다. 이와 같이 연계데이터를 관리하는데 있어서 고려해야 할 사항은 다음 〈그림 5〉와 같다.

연계 메타데이터 관리에서 연계 기준 설정은 기관의 연계 목적 및 원칙에 따라 연계유형을 포함하는 연계 대상 데이터 송수신 방법 등에 대한 기준을 설정하는 것이며, 연계 목록 관리를 통해 연계 관리 대상 데이터 측면에서 연계하는 데이터의 유형, 소재지, 연계 유형, 연계 기간 등에 대한 목록화와 관리를 실시하는 것이 필요하다. 연계 메타데이터 스키마 관리는 데이터 송수신 등 연계 관리의 기준이 되는 연계 메타데이터에 대한 표준을 정하고 이를 준수할 수 있도록 관리하는 것을 의미하며, 출처 정보 관리는 내부연계 및 외부연계 시 출처정보를 정의하고 확인하는 것을 말한다.

데이터 흐름 및 추적관리에서 데이터 출처 관리는 연계항목을 제공하는 데이터베이스, 테이블, 컬럼 등에 대한 출처 정보를 확인 및 관리하는 것을 의미하며 데이터 흐름 관리를 통해 데이터 출처 정보의 정합성 진단을 실시하고 이를 바탕으로 누락 및 오류 진단을 수행할 수 있다. 또한 데이터 추적 정보 제공 및 관리를 통해 데이터 출처 및 활용처의 정보를 제공 및 관리할 수 있도록 해야 한다.

데이터 송수신 관리에서 방법 관리는 기관 간 협의를 통한 연계 계획 수립 및 데이터 송수신 방법을 관리하는 것을 포함하는데 예를 들어 OAI-PMH 방법을 사용할 것인지, OpenAPI를 활용할 것인지, 혹은 파일을 직접 수집할 것인지에 대한 방법을 관리하는 것을 의미한다. 기간 설정은 기관 간 협의를 통한 연계 기간 뿐만 아니라 데이터 송수신 기간 및 시기를 설정하는 것을 의미하는데 이는 실시간으로 업데이트를 반영할 것인지 혹은 주기적으로 업데이트를 반영할 것인지 등에 대한 내용을 포함할 수 있다. 끝으로 송수신 장애 및 오류 관리는 오류 분석 및 개선 계획 수립 후 개선 조치를 수행하



〈그림 5〉 연계데이터 품질관리

는 것을 의미한다.

연계데이터 뿐만 아니라 연구데이터 서비스 플랫폼을 통해 직접 등록된 데이터는 모두 수집기를 통해 수집DB에 저장하게 되는데, 이렇게 저장된 데이터는 관리DB로 이관된다. 관리DB로 이관되는 과정에서 데이터 매핑과 품질 검증이 진행되어야 하며, 데이터 품질 검증을 위해 데이터 품질기준을 구축 및 활용하게 된다. 연구데이터 품질기준은 연구데이터의 품질 수준을 측정하기 위한 관점을 정의한 것으로 품질 검증의 기준이 되며, 연구데이터 품질 검증을 위해 사용될 수 있는 데이터 품질기준을 다음 〈표 4〉와 같이 제시한다. 해당 품질기준은 연구데이터 서비스 플랫폼에서 수집, 관리, 제공 및 보존하는 데이터의 품질을 진단하는데 활용되어질 수 있다.

뿐만 아니라 데이터가 등록, 수집, 관리, 서비스되는 전 과정에서 데이터 표준화, 큐레이션, 품질진단, 흐름관리, 운영관리, 산출물관리, 라이선스 정책 관리, 사용자관리 등이 수행되어야 한다(〈그림 4〉 참조).

데이터 표준화는 시스템별로 산재해 있는 데이터 정보요소에 대한 명칭, 정의, 형식, 규칙에 대한 원칙을 수립하는 것을 의미하며, 데이터 큐레이션은 데이터 생애주기에 따라 디지털 데이터를 관리함으로써 아카이빙뿐만 아니라 검색 및 향후 재사용 가능성을 위한 데이터 관리 및 그에 관련한 모든 활동을 포함하는 개념이다. 데이터 품질진단을 통해 수립된 데이터 품질기준을 바탕으로 데이터의 품질 수준을 진단 및 평가하는 과정이 필요하며, 데이터 흐름관리를 실시함으로써 데이터의 생산부터 보존까지 단계별

〈표 4〉 연구데이터 품질기준

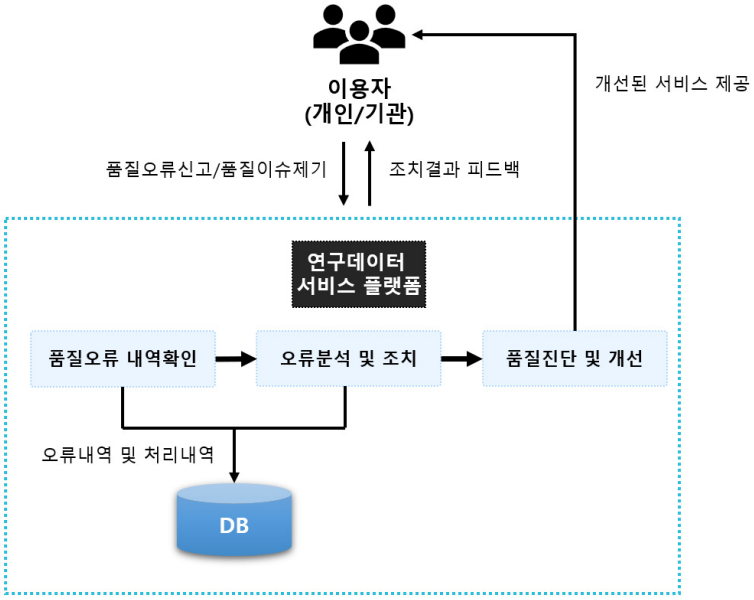
지표	정의
유효성 (Validity)	데이터 항목은 정해진 데이터 유효범위 및 도메인을 충족해야 함
완전성 (Completeness)	필수항목에 누락이 없어야 함
일관성 (Consistency)	데이터가 지켜야 할 구조, 값, 표현되는 형태가 일관되게 정의되고, 서로 일치해야 함
접근성 (Accessibility)	사용자가 원하는 데이터에 손쉽게 접근하여 사용할 수 있어야 함
정확성 (Accuracy)	데이터 입력 단계부터 오류가 없도록 하고, 저장된 데이터가 유효한 값으로 존재해야하며, 실세계에 존재하는 객체의 표현 값으로 정확히 반영이 되어야 함
고유성 (Uniqueness)	데이터 항목은 유일해야 하며 중복되어서는 안 됨
적시성 (Timeliness)	정보의 발생, 수집, 그리고 갱신 주기를 유지해야 함
유용성 (Usefulness)	사용자가 만족하는 수준의 충분한 정보가 수집 및 제공되고, 사용자의 정보 이용에 만족도를 충족시키며, 데이터의 범위와 상세화 정도를 충족시켜야 함
보안성 (Security)	데이터 관리 주체가 관리되고 외부 및 내부 요인으로부터 데이터를 보호하기 위해 접근이 적절히 통제되며 개인정보 등 주요 데이터에 대하여 보호 조치가 이루어져야 함

활용 및 출처를 관리하여 흐름 정보를 제공할 수 있도록 해야 한다. 데이터 운영관리는 기관에서 정의한 품질관리 정책, 지침에 따른 데이터 관리 프로세스의 원활한 수행이 가능하도록 담당 역할을 정의하고 운영하는 것을 말하며, 데이터 산출물관리는 시스템 구축 단계에서 생성되는 산출물들 가운데 데이터 품질과 관련된 산출물을 생성 및 검증하고 현재까지의 진행 사항을 반영하는 일련의 활동이라고 이해할 수 있다. 데이터 라이선스 정책 관리를 실시함으로써 데이터의 공개 혹은 공유 시 적절한 라이선스를 선정하고 적용하여 저작자의 저작권을 보호할 수 있도록 정책 및 지침을 구축하고 관리하여야 하며, 데이터 사용자관리를 통해 데이터를 이용하는 사용자 관리, 접속현황 및 이력관리, 사용자의 접근권한 관리 등 데이터의 생산자 및 이용자를 관리하고 지원할 수 있어야 한다.

4.3 활용단계 품질관리

활용단계 품질관리는 기관이 생성 및 보유하고 있는 데이터를 활용함에 있어서 발생할 수 있는 품질 이슈를 인지하여 개선하는 일련의 활동을 포함하는 개념으로 이해할 수 있다(한국정보화진흥원, 2018). 제공 데이터의 품질을 유지하기 위해서는 데이터를 이용하는 이용자가 데이터를 활용하는 과정에서 발견하는 다양한 오류를 관리 기관이 알 수 있도록 기록하고 관리하는 것이 요구되는데, 활용단계 품질관리에는 개인 및 기관을 포함하여 데이터를 이용하는 이용자로부터 품질 오류에 대한 신고를 접수 받아 확인·조치 및 개선하는 과정으로 구성된 품질오류 신고 관리가 포함되며, 이를 도식화하면 다음 〈그림 6〉과 같다.

개인과 기관을 포함하는 개념인 이용자가 데



〈그림 6〉 품질오류신고 관리

이터를 활용하는 과정에서 오류를 발견하면, 해당 내용에 대한 품질오류신고 또는 품질이슈 제기를 데이터 제공 플랫폼 또는 담당 기관에 전달하게 되고, 기관의 데이터 품질오류 담당자는 오류신고 내용을 파악한 후 오류내역에 대한 분석을 진행하고 담당자에게 이관하게 된다. 품질관리 실무담당자는 해당 오류를 분석하고 처리하며, 이 과정에서 오류내역 및 처리내역은 DB에 저장하도록 한다. 오류처리 후 개선사항에 대해서는 신고자에게는 통지할 수 있도록 하고, 단기간 내에 오류내역을 처리할 수 없을 경우에는 오류처리 계획에 대한 내용을 신고자에게 알림으로써 지속적인 모니터링이 가능케 할 필요가 있다.

5. 결론 및 제언

데이터의 양적 증가와 더불어 질적 관리에 대한 중요성 역시 주목받고 있으며, 연구데이터를 대상으로 한 공유, 활용, 관리에 대한 중요성이 증대함과 동시에 품질관리 역시 같은 맥락에서 주목받고 있다. 본 연구는 여러 품질관리 모델을 조사하여 비교・분석함으로써 데이터 품질관리에 필요한 공통 요인들을 추출하였고, 이를 바탕으로 연구데이터 특성에 맞는 품질관리 프로세스 모델을 제안하였다. 특히 본 연구는 연구데이터를 대상으로 하여 수집에서 서비스까지 일련의 체계에서 품질관리가 어떻게 진행될 수 있는지에 대한 논의를 진행했다는 점에서 의의를 갖는다.

본 연구는 공공데이터 품질관리 모델, 빅데이터 품질관리 모델, 그리고 연구데이터 관리를 위한 데이터 생애주기 모델을 분석하여 각 품질관리 모델에서 공통적으로 나타나는 구성요인을 분석하였다. 품질관리 모델은 품질관리를 수행하는 객체인 대상 데이터의 특성에 따라서 생애주기에 맞추어 혹은 PDCA 모델을 바탕으로 구축되고 제안되었는데 공통적으로 계획, 수집 및 구축, 운영 및 활용, 보존 및 폐기의 구성요소가 포함됨을 파악하였다.

이를 바탕으로 본 연구는 연구데이터를 대상으로 한 품질관리 프로세스 모델을 제안하였다. 제안된 연구데이터 품질관리 프로세스 모델은 계획, 구축 및 운영, 활용단계로 구성되어 있으며 제안된 내용을 간략히 정리하면 다음과 같다.

첫째, 계획단계 품질관리는 연구데이터의 품질관리 계획을 수립하는 단계에서 고려해야 하는 요인을 의미하며, 해당 단계에서는 품질진단 및 개선계획, 표준화 적용계획, 연계데이터 품질 확보 계획, 품질관리 산출물 점검 및 최신성 확보 계획, 품질관리 인프라 구축 및 운영계획, 윤리 규범 및 사용 지침 관리 등을 고려하고 수행해야 한다.

둘째로, 구축 및 운영단계의 품질관리에서는 데이터 표준화, 데이터 산출물 관리, 연계데이터에 대한 품질관리, 그리고 품질관리 계획 수립단계에서 선정된 데이터베이스 및 데이터에 대한 품질진단 및 개선활동을 수행해야 한다. 또한 해당 단계에서는 데이터가 수집되어 연구데이터 서비스 플랫폼에서 서비스되기까지 각 DB에서 DB로 이동하는 단계마다 특성에 맞는 품질관리 활동이 진행되어야 한다.

마지막으로 활용단계 품질관리는 기관이 생성 및 보유하고 있는 데이터를 활용하는데 있어서 발생할 수 있는 품질 이슈를 인지하고 개

선하는 일련의 활동을 포함하는 개념이며, 여기에는 품질오류 신고관리가 포함된다.

최근 특히나 공공데이터를 중심으로 품질관리에 대한 논의가 활발하게 진행되었는데, 연구데이터는 해당 데이터가 갖는 중요도에 비해 품질관리에 대한 논의는 부족한 실정이다. 본 연구는 연구데이터를 대상으로 하여 품질관리를 수행하는 보다 체계적인 모델을 제안했다는 것에 의의를 갖으며 해당 모델이 향후 연구데이터 품질관리 활동에 도움이 되기를 기대하는 바이다. 특히나 본 연구에서 제안되는 연구데이터 품질관리 프로세스 모델은 실제 연구데이터 서비스 플랫폼에서 데이터를 수집하여 서비스하는 일련의 과정에 맞추어 제안된 모델로서, 이론적인 개념을 제안하는 것뿐만 아니라 시스템적으로 각 단계에 맞추어 직접적인 적용이 가능할 것이라는 기대를 갖는다. 예를 들어, 국내의 대표적인 연구데이터 서비스 플랫폼이라고 할 수 있는 국가연구데이터플랫폼인 DataON과 같은 곳에서부터 해당 품질관리 모델을 발전 및 적용시킬 수 있을 것으로 기대하며, 궁극적으로는 연구데이터를 서비스함에 있어서 데이터의 품질을 제고하여 이용자들의 만족도를 충족시킬 수 있는 방안이 될 수 있을 것이다. 그러나 동시에 모든 연구데이터 서비스 플랫폼이 동일한 시스템과 DB 흐름을 갖고 있고 운영되는 것이 아니기 때문에 각 시스템 운영의 차이에 따라 적용에 어려움이 발생할 수 있다는 한계점 역시 존재한다.

참 고 문 헌

과학기술기본법. 법률 제18727호.

국가과학기술연구회 (2019). 연구데이터 관리 가이드라인 (2019-07).

국가연구개발혁신법. 법률 제18645호.

김선호, 이창수(2013). 데이터 품질관리 프로세스 평가를 위한 프로세스 참조모델. 한국전자거래학회지, 18(4), 83-105. https://doi.org/10.7838/jsebs.2013.18.4.083

김주섭, 김선태, 전예린 (2019). 연구 데이터 관리를 위한 데이터 라이프 사이클 제안. 한국문헌정보학회지, 53(4), 309-340. https://doi.org/10.4275/KSLIS.2019.53.4.309

김형섭 (2020). 데이터 품질관리 평가 모델에 관한 연구. 한국융합학회논문지, 11(7), 217-222. https://doi.org/10.15207/JKCS.2020.11.7.217

박고은, 김창재 (2015). 공공개방데이터 품질 특성에 관한 연구. 디지털융복합연구, 13(10), 135-146. https://doi.org/10.14400/JDC.2015.13.10.135

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

안전행정부 (2014). 공공데이터 관리지침, 제2014-13호.
정혜정 (2007). 데이터 품질 평가에 관한 연구. 인터넷정보학회논문지, 8(4), 119-128.
한국과학기술정보연구원 (2019). 연구데이터 공유확산체제 구축(K-19-L01-C03).
한국데이터베이스진흥원 (2006). 데이터 품질관리 지침(Ver 2.1).
한국정보통신기술협회 (2022). Verifiable Credentials Data Model 1.1.
한국정보화진흥원 (2015). 공공데이터 품질관리 수준평가 모델(안). 한국정보화진흥원 제13차 개방품질 전문위원회 보고자료.
한국정보화진흥원 (2018). 공공데이터 품질관리 매뉴얼 v2.0.
한국지능정보사회진흥원 (2021). 빅데이터 플랫폼 및 센터 데이터 품질관리 가이드.
한나은, 김성희 (2014). 외국 대학도서관의 디지털 큐레이션 프로세스 비교분석. 한국도서관・정보학회지, 45, 93-116. https://doi.org/10.16981/kliss.45.2.201406.93
Data quality - Part 1: Overview. ISO 8000-1:2022.
Eckerson, W. (2002). Data warehousing special report: Data quality and the bottom line. Applications Development Trends, 1(1), 1-9.
English, L. P. (2009). Information quality applied: Best practices for improving business information, processes and systems. New Jersey: Wiley.
Kindling, M. & Strecker, D. (2022). Data Quality Assurance at Research Data Repositories. Data Science Journal, 21(1). http://doi.org/10.5334/dsj-2022-018
National Science Foundation (2014). Proposal and award policies and procedures guide (nsf15001).
Wang, R. Y., Ziad, M., & Lee, Y. W. (2006). Data quality. Vol. 23. Berlin: Springer Science & Business Media.

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

Framework Act on Science and Technology. No. 18727.
Han, Na-Eun & Kim, Seong-Hee (2014). Comparative analysis on digital curation process in foreign academic libraries. The Korea Journal of Library and Information Science, 45, 93-116. https://doi.org/10.16981/kliss.45.2.201406.93
Jung, Hye-Jung (2007). A study of the data quality evaluation. Journal of Internet Computing and Services, 8(4), 119-128.
Kim, Hyung-Sub (2020). A study on the data quality management evaluation model. Journal of the Korea Convergence Society, 11(7), 217-222.

https://doi.org/10.15207/JKCS.2020.11.7.217

Kim, Juseop, Kim, Suntae, & Jeon Yerin (2019). Data life cycle proposal for research data management. Journal of the Korean Society for Library and Information Science, 53(4), 309-340. https://doi.org/10.4275/KSLIS.2019.53.4.309

Kim, Sunho & Lee, Changsoo (2013). The process reference model for the data quality management process assessment. The Journal of Society for e-Business Studies, 18(4), 83-105. https://doi.org/10.7838/jsebs.2013.18.4.083

Korea Data Agency (2006). Data Quality Management Guidelines(Ver 2.1).

Korea Institute of Science and Technology Information (2019). Establishment of Research Data Sharing and Dissemination System (K-19-L01-C03).

Ministry of Security and Public Administration (2014). Government Data Management Guidelines, No. 2014-13.

National Information Society Agency (2015). Government Data Quality Management Level Evaluation Model. Report data of the 13th Open Quality Expert Committee of the National Information Society Agency.

National Information Society Agency (2018). Open Government Data Quality Management Manual v2.0.

National Information Society Agency (2021). Big Data Platform and Center Data Quality Management Guide.

National Research and Development Innovation Act. No. 18645.

National Research Council of Science and Technology (2019). Research Data Management Guidelines (2019-07).

Park, Go-Eun & Kim, Chang-Jae (2015). Quality characteristics of public open data. Journal of Digital Convergence, 13(10), 135-146. https://doi.org/10.14400/JDC.2015.13.10.135

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

Telecommunications Technology Association (2022). Verifiable Credentials Data Model 1.1.