# CoreTrustSeal 인증 획득을 통한 데이터 리포지토리의 신뢰성 향상을 위한 연구*

## A Study to Improve the Trustworthiness of Data Repositories by Obtaining CoreTrustSeal Certification

이혜림 (Hea Lim Rhee)** 엄정호 (Jung-Ho Um)***
신영호 (Youngho Shin)**** 임형준 (Hyung-jun Yim)*****
한나은 (Na-eun Han)******

## 초 록

데이터의 가치에 대한 인식이 높아지면서 데이터를 관리, 보존, 활용하는 데 있어서 데이터 리포지토리의 역할이 점점 더 중요해지고 있다. 본 연구에서는 CoreTrustSeal(CTS) 인증 획득을 한 리포지토리의 신청서를 비교분석하여, 데이터 리포지토리의 신뢰성을 높이는 방법을 조사한다. 데이터 리포지토리에 대한 신뢰는 데이터 보호뿐만 아니라 리포지토리와 이해관계자 간의 신뢰를 구축하고 유지하는 데에도 중요하며, 이는 결과적으로 데이터 보존 및 활용에 대한 연구자의 결정에 영향을 미친다. 먼저, 본 연구에서는 신뢰할 수 있는 데이터 리포지토리에 대한 국제 인증인 CTS를 조사하여 리포지토리의 신뢰성과 효율성에 미치는 영향을 분석한다. 그리고 한국과학기술정보연구원(KISTI)이 운영하는 국내 최초 CTS 인증 리포지토리인 DataON을 사례로 CTS 인증을 획득한 4개 리포지토리를 비교 분석한다. 여기에는 DataON, NASA의 PO.DAAC, 제네바 대학의 Yareta 및 독일의 DARIAH-DE 리포지토리가 포함된다. 본 연구에서는 이러한 리포지토리가 CTS가 정한 필수 요구 사항을 어떻게 충족하는지 조사하고, 데이터 리포지토리의 신뢰성을 향상시키기 위한 전략을 제안한다. 주요 조사 결과에 따르면, CTS 인증을 획득하려면 데이터 리포지토리는 조직 인프라, 디지털 객체 관리 및 기술 측면에서 정책, 시스템, 자원 관리 등을 체계적이고 효율적으로 수행하고 있고, 이를 CTS인증서에 명확하게 서술하고 근거를 보여주어야 한다. 본 연구는 투명한 데이터 프로세스, 강력한 데이터 품질 보증, 향상된 접근성 및 유용성, 지속 가능성, 보안 조치, 법적 및 윤리적 표준 준수의 중요성을 강조한다. 이러한 전략을 구현함으로써 데이터 저장소는 신뢰성과 효율성을 향상시킬 수 있으며 궁극적으로 과학 분야에서 더 폭넓은 데이터 공유 및 활용을 촉진할 수 있다.

## ABSTRACT

As the recognition of data's value increases, the role of data repositories in managing, preserving, and utilizing data is becoming increasingly important. This study investigates ways to enhance the trustworthiness of data repositories through obtaining CoreTrustSeal (CTS) certification. Trust in data repositories is critical not only for data protection but also for building and maintaining trust between the repository and stakeholders, which in turn affects researchers' decisions on depositing and utilizing data. The study examines the CoreTrustSeal, an international certification for trustworthy data repositories, analyzing its impact on the trustworthiness and efficiency of repositories. Using the example of DataON, Korea's first CTS-certified repository operated by the Korea Institute of Science and Technology Information (KISTI), the study compares and analyzes four repositories that have obtained CTS certification. These include DataON, the Physical Oceanography Distributed Active Archive Center (PO.DAAC) from NASA, Yareta from the University of Geneva, and the DARIAH-DE repository from Germany. The research assesses how these repositories meet the mandatory requirements set by CTS and proposes strategies for improving the trustworthiness of data repositories. Key findings indicate that obtaining CTS certification involves rigorous evaluation of organizational infrastructure, digital object management, and technological aspects. The study highlights the importance of transparent data processes, robust data quality assurance, enhanced accessibility and usability, sustainability, security measures, and compliance with legal and ethical standards. By implementing these strategies, data repositories can enhance their reliability and efficiency, ultimately promoting wider data sharing and utilization in the scientific community.

키워드: 데이터 리포지토리, 신뢰, 인증, CoreTrustSeal, CTS
data repository, trust, certification, CoreTrustSeal, CTS

---

* 이 논문은 2024년도 한국과학기술정보연구원(KISTI)의 기본사업으로 수행된 연구입니다. (과제번호: (KISTI)K24L1M2C3)
** 한국과학기술정보연구원 책임연구원(rhee.healim@kisti.re.kr) (제1저자)
*** 한국과학기술정보연구원 책임연구원(jhum@kisti.re.kr) (공동저자)
**** 한국과학기술정보연구원 책임연구원(shinyh@kisti.re.kr) (공동저자)
***** 한국과학기술정보연구원 선임연구원(hjyim@kisti.re.kr) (공동저자)
****** 한국과학기술정보연구원 박사후연구원(betterhan@kisti.re.kr) (교신저자)
■ 논문접수일자 : 2024년 5월 16일 ■ 최초심사일자 : 2024년 5월 31일 ■ 게재확정일자 : 2024년 6월 8일
■ 정보관리학회지, 41(2), 245-268, 2024. http://dx.doi.org/10.3743/KOSIM.2024.41.2.245

# 1. 서 론

데이터의 가치에 대한 인식이 높아지면서 데이터를 관리, 보존, 이용할 수 있게 해주는 데이터 리포지토리의 역할이 점점 중요해지고 있다. 데이터 리포지토리의 신뢰성은 데이터 보호의 기본일 뿐만 아니라 데이터 리포지토리와 이해관계자 간의 신뢰를 구축하고 유지하는데 중추적인 역할을 하고, 연구자들이 리포지토리의 기탁과 이용 여부를 결정하는 필수 요소로 작용한다(Azeroual & Schöpfel, 2021; Lin et al., 2020). 또한, 리포지토리의 신뢰성은 데이터 생산자들이 그들의 데이터를 기탁하도록 하는 하나의 요인으로 작용할 수 있으며, 잠재적인 데이터 이용자들에게는 그들이 접근 및 이용하는 데이터의 신뢰성과 무결성을 보장한다. 이처럼, 신뢰할 수 있는 데이터 리포지토리는 해당 리포지토리가 소장하고 있는 데이터의 품질을 보장함으로써, 해당 리포지토리의 효율성을 촉진하는데 중요한 요인으로 평가된다(Frank et al., 2017). 또한, 데이터 리포지토리는 보유 데이터에 대한 장기적인 접근 및 이용을 보장하여야 하는데, 이는 지정 커뮤니티(designated community)라고 불리우는 해당 리포지토리 데이터의 주요한 생산자 혹은 이용자와의 적극적인 소통을 기반으로 해야 한다.

데이터 리포지토리가 해당 리포지토리의 신뢰성을 확보함과 동시에 보유하고 있는 데이터의 신뢰성을 입증하기 위한 다양한 활동이 존재하는데, 이는 내부적인 관리 방안에서부터 외부의 평가를 통한 인증을 획득하는 것까지 다양하다. 외부적으로 리포지토리의 품질 수준에 대한 객관적인 평가를 받고 공인된 인증을 획득하는 것은 해당 리포지토리 보유 데이터의 신뢰성을 입증하는 주요한 활동 중 하나라고 볼 수 있다(Gualo et al., 2021). 다양한 리포지토리가 보유 데이터 및 리포지토리 자체의 신뢰성을 확보하기 위한 방법 중 하나로 CoreTrustSeal(이하 CTS)와 같은 신뢰할 수 있는 데이터 리포지토리에게 주어지는 인증을 검토하고 있으나, 해당 인증과 같은 내용이 리포지토리를 사용하는 주요 이용자 및 데이터 생산자에게 얼마나 영향을 미치는지는 아직 명확하게 밝혀진 바가 없다. 그러나 CTS로 대표되는 신뢰할 수 있는 리포지토리에게 주어지는 인증은 해당 리포지토리가 데이터를 수집, 관리, 보존, 이용하는 일련의 과정에서 구축된 표준을 얼마나 준수하고 있는지 등을 포함한 다양한 외부 검증을 실시한다. 그렇기 때문에 신뢰성에 대한 인증을 획득한 것 자체가 해당 리포지토리를 이용하는 다양한 이용자들에게 직접적인 영향을 미치지 못한다고 하더라도, CTS와 같은 외부 인증을 받은 과정을 거치면서 리포지토리는 보유한 데이터, 리포지토리 운영 자체 등에 대한 점검과 보완 과정을 통해 리포지토리의 신뢰성을 확보하기 위한 기초적인 토대를 갖출 수 있다(Corrado, 2019, 61).

CTS는 지속 가능하고 신뢰할 수 있는 데이터 인프라 구축을 장려하기 위해 다양한 분야의 데이터 리포지토리에 핵심 인증(core certification)을 부여하는 국제적인 비영리 조직이다(CoreTrustSeal, 2024). CTS에서 요구하는 'CoreTrustSeal Trustworthy Data Repositories Requirements'는 크게 조직 인프라, 디지털 객체 관리, 기술(technology)적 측면에서 데이터 리포지토리가 신뢰성을 확보하기 위해 필요한

다양한 요구 사항을 구현할 것을 명시하고 있다. 리포지토리가 신뢰성을 인정받고 CTS 인증을 획득하기 위해서는 각각의 요구 사항에 대하여 해당 리포지토리의 현황을 자체적으로 조사 및 평가하고 해당 내용을 명확하게 온라인 신청서에 서술하여 온라인에서 제출해야 한다. 제출된 내용은 CTS에서 배정한 복수의 심사자들로부터 평가를 받음으로써 CTS 인증을 획득하거나 획득하지 못한다.

데이터 리포지토리를 대상으로 외부적인 평가를 진행하고 인증을 부여하는 제도는 CTS가 유일한 것은 아니다. 그러나 CTS는 2024년 5월 현재 국제적으로 인정받고 있는 인증 제도 중 하나이며, 리포지토리 보유 데이터의 무결성과 접근성을 보장하여 다양한 연구자 및 커뮤니티에서 안정적이고 장기적으로 데이터를 접근 및 이용할 수 있도록 지원하고 있다. 전 세계의 연구데이터 리포지토리 정보를 제공하는 re3data.org에 등록된 리포지토리 2,800여 개 가운데 국제적 인증을 받은 리포지토리는 213개이며, 이 가운데 CTS 인증을 획득한 리포지토리는 116개로 총 54.5%이다. 이는 타 인증과 비교했을 때 과반이 넘는 가장 높은 비율을 차지하고 있는 것을 알 수 있다(김주섭, 양성준, 김선태, 2022). CTS 인증을 획득한다는 것은 리포지토리가 핵심적으로 갖추어야 할 기능과 정책을 구축하고 있는 시스템임을 객관적 기준을 바탕으로 한 외부 평가에서인정받았다는 것을 의미한다. 그러나 2024년 5월 기준, 전 세계적으로 다양한 국가의 125개의 리포지토리가 CTS 인증을 획득하였음에도 불구하고 국내의 CTS 인증을 받은 리포지토리는 한국과학기술정보연구원(Korea Institute of Science and Technology Information, 이하 KISTI)이 운영하는 DataON 뿐이다.

KISTI가 운영하는 DataON은 2018년에 발표된 '연구데이터 공유 이용 전략(안)'에 따라 국가연구데이터를 수집하고 검색할 수 있도록 개발된 국가연구데이터플랫폼이며, 연구자들의 개별적 데이터 기탁뿐만 아니라 국내외의 다양한 리포지토리와의 연계를 통해 광범위한 과학기술 분야 데이터를 제공하는 플랫폼이다. 뿐만 아니라 연구자가 DataON을 이용하여 데이터 융합 및 분석을 수행할 수 있도록 다양한 분석도구 및 가상작업환경 역시 제공하고 있는 원스톱 통합플랫폼을 지향한다. 과학기술 분야 전반의 연구데이터의 수집에서부터 이용에 이르기까지 광범위한 서비스를 제공하는 국가연구데이터플랫폼으로서 DataON은 다양한 지정 커뮤니티 및 연구자들로부터 신뢰성을 확보하기 위한 방안으로 2022년부터 CTS 인증을 준비하였다. 외부의 평가 및 내부적인 발전 방향의 구축을 기반으로 인증 신청을 시작하여, 2023년부터는 DataON을 직접 관리하는 KISTI 내부 전문가들로 팀을 구성하여 거버넌스를 구축하고 리포지토리를 개선하였다. 그 결과로 KISTI의 DataON이 2024년 국내에서는 최초로 CTS 인증을 획득함으로써 신뢰할만한 데이터 리포지토리로 인정받았다.

본 연구는 국내 최초로 CTS 인증을 획득한 KISTI의 DataON을 포함하여 총 4개의 데이터 리포지토리를 사례로 선정하여 CTS에서 명시하고 있는 리포지토리의 신뢰성 확보를 위해 요구되는 필수 요건들을 기준으로 삼아 각 리포지토리의 구현 내용을 비교 분석하였다. 이를 바탕으로 데이터 리포지토리의 신뢰성 향상

을 위한 방안을 제안한다. 본 연구에서 분석한 CTS 인증을 받은 여러 나라의 리포지토리의 구현 내용은 특정 리포지토리가 그들이 보유한 데이터 및 리포지토리 자체의 신뢰성을 확보 및 향상시키기 위한 방안을 모색하는데 참고자료로 이용할 수 있을 것으로 기대된다.

## 2. 데이터 리포지토리의 신뢰성

### 2.1 데이터 리포지토리에서 신뢰의 중요성

현대 데이터 중심의 환경에서 데이터 리포지토리는 연구, 정책 수립, 비즈니스 인텔리전스를 촉진하는 데이터 관리에 있어서 중요한 역할을 한다. 이러한 리포지토리에 대한 신뢰는 데이터의 신뢰성과 무결성을 뒷받침하고, 이후 그 데이터의 모든 사용에 영향을 미치기 때문에 매우 중요하다.

데이터 리포지토리에 대한 신뢰는 데이터 품질과 무결성을 보장하는 데에서부터 시작된다. 좋은 데이터 리포지토리는 엄격하게 데이터 관리를 실행하여 데이터 세트가 정확하고 완전하며 신뢰할 수 있도록 보장한다(Oliver & Harvey, 2016). 여기에는 시간이 지나도 데이터의 정확성과 관련성을 유지하기 위한 강력한 검증 프로세스, 정기적인 감사, 그리고 업데이트가 포함된다. 연구자와 데이터 분석가가 데이터의 무결성을 신뢰할 때, 해당 정보가 철저하게 관리되었다고 생각하고, 그 정보를 바탕으로 자신의 연구 결과와 결정을 내릴 수 있다(Azeroual & Schöpfel, 2021; Yoon, 2014).

데이터 리포지토리에 대한 신뢰는 해당 리포지토리가 보유하고 있는 데이터로의 접근가능성과 이용가능성을 향상시킨다(Trisovic et al., 2021). 향상된 접근가능성과 이용가능성은 데이터의 광범위한 이용을 장려하고 오픈 사이언스와 공동 연구 환경을 조성한다(Downs, 2021; Xafis & Labude, 2019). 신뢰할 수 있는 리포지토리는 데이터를 이용할 수 있을 뿐만 아니라 다양한 수준의 전문 지식을 가진 이용자가 쉽게 접근할 수 있도록 보장한다. 여기에는 포괄적인 메타데이터, 명확한 문서화, 이용자 친화적인 인터페이스를 제공하는 것이 포함된다.

윤리적 지침과 법적 표준을 준수하는 것은 데이터 리포지토리에 대한 신뢰를 구축할 수 있게 만드는 또 하나의 기반이 된다. 리포지토리는 데이터 수집과 배포를 합법적이고 책임감 있게 수행하면서, 동시에 개인 정보 보호법 및 윤리 규범에 따라 민감한 데이터를 관리해야 한다(Broekstra, 2021; Paprica et al., 2023). 리포지토리가 기밀성을 존중하고 규제 요구 사항을 준수하여 이용자와 데이터 생산자의 이익을 보호한다는 사실을 이용자가 알게 될 때, 해당 리포지토리에 대한 이용자의 신뢰가 증가한다.

학술 및 과학 커뮤니티에서는 데이터 리포지토리에 대한 신뢰가 기본이다. 연구자들은 그들의 포괄적이고 인용 가능한 데이터 세트를 제공하기 위해 신뢰할 수 있는 리포지토리에 의존한다(Broekstra, 2021; Downs, 2021; Yoon, 2014). 이러한 신뢰는 연구 결과의 재현성과 검증을 촉진하며, 이는 하나의 학문 내에서 그리고 학문간 협력하는 융합연구에서 지식을 발전시키는 데 중요한 역할을 한다. 따라서 신뢰할 수 있는 리포지토리는 연구의 무결성과 연구의

진행을 지원하면서, 과학 인프라의 중추적인 토대가 된다.

## 2.2 신뢰할 수 있는 데이터 리포지토리의 조건

신뢰할 수 있는 데이터 리포지토리는 이용자 커뮤니티에 효과적으로 서비스를 제공하고 관리하는 데이터의 무결성을 유지하기 위해 여러 가지 요건을 충족해야 하는데, 그 중 주요한 요건들로는 다음과 같은 것들이 언급되고 있다.

### 2.2.1 투명성

데이터 처리 프로세스, 데이터 원본, 데이터 원본의 수정 사항에 관한 투명성은 필수적이다. 신뢰할 수 있는 리포지토리는 데이터를 수집한 사람, 수집 방법, 시간 경과에 따른 데이터 변경 사항 등 데이터 출처에 대한 명확한 문서를 제공한다. 이러한 투명성은 이용자가 자신의 요구에 대한 데이터의 적합성을 평가할 수 있도록 하는 데 중요하다(Azeroual & Schöpfel, 2021; Donaldson, 2020; Frank et al., 2017; Lin et al., 2020).

### 2.2.2 데이터 품질 보장

신뢰할 수 있는 리포지토리는 데이터 품질을 보장하기 위해 엄격한 절차를 실행한다. 그러한 절차로는 데이터의 정확성을 검토하고 접근 가능 하게 만들기 위한 데이터 큐레이션, 오류 검사, 검증 프로세스를 위한 메커니즘이 있다. 일관된 데이터 품질 보장은 데이터 세트의 신뢰성을 유지하는 데 도움이 된다(Liaw et al., 2021; Stvilia & Lee, 2024).

### 2.2.3 접근가능성과 이용가능성

다양한 수준의 전문 지식과 다양한 학문적 배경을 가진 이용자들이 리포지토리의 데이터에 접근할 수 있어야 한다(Lin et al., 2020; Volentine et al., 2015). 여기에는 직관적인 이용자 인터페이스, 효과적인 검색 도구, 데이터를 자세히 설명하는 포괄적인 메타데이터 제공이 포함된다. 또한, 리포지토리는 데이터를 효과적으로 사용하는 방법에 대한 지원과 지침을 제공해야 한다(Prieto, 2009).

### 2.2.4 지속가능성

신뢰할 수 있는 리포지토리는 장기적인 데이터 보존과 접근을 보장하기 위해서 기관의 지속가능성을 보장하는 모델이 필요하다(Yoon, 2014). 그러한 모델에는 명확한 비즈니스 모델, 적절한 자금 조달, 그리고 최신 기술로의 데이터 마이그레이션 계획이 포함된다. 리포지토리의 지속성을 보장하는 것은 연구자들의 지속적인 연구와 장기 연구를 지원하는 데 있어 중요하다(Lin et al., 2020).

### 2.2.5 보안 조치

무단 접근, 위반, 기타 위협으로부터 데이터를 보호하려면 강력한 보안 조치가 필요하다. 여기에는 보안 저장소, 데이터 암호화, 액세스 제어, 정기적인 보안 감사 실행이 포함된다(Broekstra, 2021; Li, 2004). 데이터 보안을 보장하면 이용자 신뢰를 구축하는 데 도움이 되며 민감한 정보를 보호할 수 있다(Edmunds et al., 2016).

### 2.2.6 법과 윤리 준수

리포지토리는 데이터 보호법(예: 유럽 연합

의 General Data Protection Regulation) 및 산업별 규정과 같은 해당 법률 및 규제 프레임워크를 준수해야 한다(European Union, 2024). 특히 개인 식별 정보나 민감한 데이터를 다룰 때, 이용자는 윤리적 표준을 준수하여 책임감을 가지고 데이터를 윤리적으로 사용해야 한다(Lin et al., 2020; Mutula, 2011).

#### 2.2.7 지역 사회 참여

커뮤니티의 요구 사항을 이해하고 피드백을 수집하기 위해 커뮤니티에 참여하는 것은 리포지토리의 지속적인 개선을 위해 필요하다. 신뢰할 수 있는 리포지토리는 그 기관의 서비스를 향상시키고 커뮤니티의 변화하는 요구 사항을 충족시키기 위해 커뮤니티의 의견을 적극적으로 요청하고 수용한다(Yakel et al., 2013; Yoon, 2014).

## 3. CTS 사례 대상과 비교분석 방법

본 연구는 CTS 인증을 획득한 국내외 데이터 리포지토리 4개를 사례로 선정하여 CTS 인증 획득을 위해 필수적으로 요구되는 내용을 기준으로 각 리포지토리를 비교 분석하고, 이를 바탕으로 데이터 리포지토리의 신뢰성 향상을 위한 방안을 제안한다.

### 3.1 CTS 인증 요구사항

CTS 인증을 획득하기 위해서는 요구사항에 맞게 온라인 신청서를 제출해야 한다. CTS 인증 요구사항은 신뢰할 수 있는 리포지토리의 특성을 설명할 수 있는 항목으로 구성되어 있고, 독립형 항목으로 이루어진 모든 요구사항에 대하여 필수적으로 기입해야 한다. 각 기입 항목에 대하여 일부 중복되는 내용이 있을 수 있으나, 이는 데이터 리포지토리의 관리 및 운영에 있어서 모든 요소들이 분리되어 작용하지 않는 것에 기인한다. CTS는 신뢰할 수 있는 리포지토리가 갖추어야 하는 특성의 최신성을 유지하기 위해 3년마다 새로운 요구사항을 제시하고, 이미 CTS 인증을 획득한 리포지토리 역시 그 권한을 3년 동안만 유지할 수 있도록 하고 있다. 이를 갱신하여 권한을 유지하기 위해서는 추후 완전히 새로운 인증 신청을 다시 진행해야 한다. 2024년 5월 기준, CTS 데이터 리포지토리 요구사항은 2023-2025년 버전을 요구하고 있으나, 본 연구에서는 CTS 데이터 리포지토리 요구사항 2020-2022년 버전을 기준으로 인증을 획득한 기관을 사례로 선정하였다. 해당 CTS 인증 요구사항을 간략히 표로 정리하면 다음 〈표 1〉과 같다.

### 3.2 사례 대상 선정 방법

본 연구는 국내에서 최초로 CTS 인증을 획득한 DataON을 포함하여 총 4개의 국내외 데이터 리포지토리를 사례로 선정하여 비교 분석하는 다중 사례 분석 연구를 진행하였다. DataON을 중심으로 비교 분석할 사례를 선정하기 위하여 총 5개의 기준을 적용하였다.

첫째, 사례 대상의 최신성을 확보하기 위해 2024년 이후 CTS 인증을 획득한 리포지토리를 선별하였으며, 이 단계에서 DataON을 포함하

〈표 1〉 CTS 데이터 리포지토리 요구사항 2020-2022 개요

| 구분 | 항목 |
|---|---|
| | R0. 맥락(Context) |
| 조직 인프라 (Organizational Infrastructure) | R1. 미션/범위(Mission/Scope) |
| | R2. 라이선스(Licenses) |
| | R3. 접근의 지속성(Continuity of Access) |
| | R4. 기밀/윤리(Confidentiality/Ethics) |
| | R5. 조직 인프라(Organizational Infrastructure) |
| | R6. 전문가 가이드(Expert Guidance) |
| 디지털 객체 관리 (Digital Object Management) | R7. 데이터의 무결성과 진본성(Data Integrity and Authenticity) |
| | R8. 평가(Appraisal) |
| | R9. 문서화된 저장 절차(Documented Storage Procedures) |
| | R10. 보존 계획(Preservation Plan) |
| | R11. 데이터 품질(Data Quality) |
| | R12. 워크플로우(Workflows) |
| | R13. 데이터 발견 및 식별(Data Discovery and Identification) |
| | R14. 데이터 재사용(Data Reuse) |
| 기술 (Technology) | R15. 기술 인프라(Technical Infrastructure) |
| | R16. 보안(Security) |

여 총 19개의 리포지토리를 확인하였다.

둘째, 비교의 기준을 명확하게 통일하기 위해 CTS의 데이터 리포지토리 요구사항 2020-2022년 버전으로 인증을 획득한 리포지토리를 추가적으로 선별하였다. 이를 통해 2023-2025년 버전으로 인증을 획득한 7개의 리포지토리를 제외하고 총 12개의 리포지토리를 확인하였다. 2020-2022년 버전에서 2023-2025년의 버전으로 변경되면서 항목별 리포지토리 요건에 차이가 있고, 인증 구현 수준의 평가 기준이 0~4에서 0~1로 변경됨에 따라 정확한 비교를 위해 동일한 버전으로 인증을 획득한 리포지토리로 대상을 한정하였다.

셋째, 각 인증 항목별로 데이터 리포지토리가 신청한 수준과 심사자들이 평가한 수준이 동일한 리포지토리만을 선별하였다. CTS 인증은 신청 리포지토리가 각 항목별로 해당 리포지토리의 수준 (인증 구현) 수준을 자체적으로 평가하여 0~4 중에서 적합한 수준을 선택하여 기입하도록 하고 있다. 이 때, 4점으로 갈수록 구축 및 운영이 안정적으로 수행되고 있다는 의미이다. 이를 기반으로 두 명의 심사자들이 평가를 진행하면서 항목별로 해당 수준이 적절한지 확인하고, 심사자들이 생각하는 수준을 제시한다. 이미 인증을 획득한 리포지토리 가운데는 리포지토리의 자체 평가와 심사자들의 평가 수준이 일치하지 않는 경우가 존재하였는데, 이 경우에는 해당 항목의 인증 구현 수준을 이해하는 기준이 명확하지 않기 때문에 사례 대상에서 제외하였다. 해당 단계를 통해 추가적으로 5개의 리포지토리를 제외하고, DataON을 포함하여 총 7개의 리포지토리를 선별하였다.

넷째, 리포지토리의 성격을 반영하여 사례를 선정하였다. 리포지토리별로 특정한 분야의데이

터로 한정하여 보유하고 있는 경우가 있고, 이와는 달리 전반적인 데이터를 아우르는 리포지토리가 있다. 비교 분석의 중심이 되는 DataON의 경우 과학기술 전분야의 데이터를 폭넓게 아우르고 있기 때문에 원활한 비교를 위해 특정 분야 중점의 리포지토리와 전반적인 데이터를 아우르는 리포지토리를 동시에 사례 대상으로 선별하였으며, 추가로 학술논문 중심의 리포지토리는 데이터 리포지토리의 특성과 비교하기에 적합하지 않아 제외하였다.

마지막으로, 각 항목별 인증 요구사항을 보다 명확하게 확인하기 위해서 최대한 각 기관별로 항목별 수준이 상이한 리포지토리를 선정하였다.

위의 5개의 기준을 적용하여 선정한 본 연구의 사례는 KISTI의 DataON, 미국 NASA Jet Propulsion Laboratory의 Physical Oceanography Distributed Active Archive Center(이하 PO.DAAC), 스위스 제네바 대학의 Yareta, 독일 Humanities Data Centre의 DARIAH-DE Repository(이하 DARIAH-DE)로 총 4개의 데이터 리포지토리이다.

### 3.3 사례 대상 리포지토리의 배경

본 연구에서 분석 대상으로 삼고 있는 4개의 리포지토리에 대해 간략하게 살펴보면 다음과 같다. 먼저 DataON은 KISTI에서 운영하는 국가연구데이터플랫폼으로, 연구자들이 연구 과정에서 생성한 데이터를 체계적으로 수집, 관리, 공유하며, 데이터의 재이용을 촉진하기 위해 사용되고 있다. DataON은 연구데이터 통합 플랫폼으로서 개인 연구자의 데이터 기탁뿐만 아니라 국내외 여러 연구 기관 및 리포지토리와의 연계를 통해 광범위한 연구데이터에 접근할 수 있도록 서비스하고 있다. DataON은 연구데이터 저장소(National Research Data Archives, 이하 NaRDA)를 포함하고 있는데, NaRDA는 타 기관에 배포가 가능한 소프트웨어로서, 이를 이용하여 타 기관과의 연계를 용이하게 할 수 있다는 이점을 가지고 있다. DataON은 국내외 여러 기관 및 리포지토리와의 연계를 진행함에 있어서, 이미 구축된 리포지토리와의 연계 뿐만 아니라 NaRDA 배포를 통한 연구데이터 관리의 컨설팅을 진행함과 동시에 시스템 연계를 용이하게 함으로써 산발적으로 분포되어 있는 연구데이터를 종합적으로 서비스하기 위해 노력하고 있다.

둘째로, PO.DAAC는 NASA의 물리해양학 분산 활성 아카이브 센터에서 제공하는 다양한 데이터와 소프트웨어를 공유하고 관리하는 플랫폼으로서, 과학자, 연구원, 일반 대중들이 해양 관측 데이터와 관련 도구들을 쉽게 탐색 및 이용할 수 있도록 지원하는 역할을 한다. PO.DAAC는 위성, 항공기, 현장 관측 등을 통해 수집된 다양한 해양 관측 데이터를 제공할 뿐만 아니라 해양 관측 데이터를 처리하고 분석하는데 도움이 되는 데이터 시각화, 통계 분석, 모델링 등을 위한 도구를 포함한 다양한 소프트웨어 도구를 제공한다. 뿐만 아니라 해양 관측 데이터와 관련된 다양한 문서와 교육 자료를 제공하는데, 이는 데이터에 대한 설명, 사용 방법, 연구 사례 등을 포함한다.

셋째로, Yareta 리포지토리는 제네바 대학교에서 운영하는 연구데이터 관리 플랫폼으로서 유럽 및 전 세계 연구자들의 연구데이터 공유

및 이용 촉진을 목적으로 구축되었다. Yareta를 통해 다양한 분야의 연구데이터를 검색 및 이용할 수 있으며, 해당 리포지토리를 통해 연구데이터의 공유 및 관리도 가능하다. Yareta는 FAIR(Findability 검색 가능성, Accessibility 접근 가능성, Interoperability 상호호환 가능성, Reusability 재사용 가능성) 원칙을 명시하여 준수하고 있으며, 다양한 데이터 형식을 지원한다. 또한 이용자들의 요구에 맞추어 서비스를 지속적으로 개선하며, 엄격한 데이터 보안 시스템을 통해 연구데이터를 안전하게 관리 및 제공할 수 있도록 노력하고 있다.

넷째로, DARIAH-DE는 독일의 인문학 및 사회과학 데이터 아카이브에서 운영하는 디지털 장기 보관소이다. 특히 인문학 및 사회과학 분야의 연구자들이 연구 과정에서 생성한 데이터를 체계적으로 관리하고 공유하며, 이용을 촉진하기 위해 구축된 리포지토리이다. DARIAH-DE를 통해 다양한 분야의 인문학 및 사회과학 연구데이터를 검색하고 다운로드하여 이용할 수 있고, 연구자들은 자신의 연구데이터를 공개적으로 공유하거나 혹은 제한적으로 공유할 수 있다. 뿐만 아니라 해당 리포지토리는 연구데이터의 생성, 처리, 분석, 보존 등을 위한 다양한 도구를 제공하며 연구데이터의 관리와 이용과 관련된 교육과 컨설팅 서비스를 제공한다. DARIAH-DE는 인문학 및 사회과학 분야의 연구에 이용되는 리포지토리이자 동시에 교육자들은 해당 리포지토리를 통해 학생들에게 인문학, 사회과학 연구데이터를 이용한 교육을 제공하며, 문화기관들은 DARIAH-DE를 통해 문화 유산 자료를 디지털화하여 보존하고 있다.

### 3.4 비교 분석 방법

본 연구에서는 CTS 인증을 획득한 총 4개의 리포지토리를 선정하고, CTS 인증 요구사항 항목별로 해당 리포지토리가 어떻게 구현되어 있는지 비교 분석하였다. CTS 인증을 획득한 리포지토리가 제출한 온라인 신청서는 각 요구사항 항목별 구현 수준을 포함하여 해당 리포지토리가 작성한 내용이 CTS 홈페이지를 통해 공개된다. 본 연구에서는 각 리포지토리가 작성하여 제출한 온라인 신청서를 기반으로, 총 16개의 인증 항목별로 필요한 요구사항을 정리하여 각 리포지토리가 해당 요구사항에 어느 수준으로 부합하는 리포지토리를 구축 및 운영하고 있는지 파악하였다. 각 인증 항목별로 해당 리포지토리의 구현 수준을 파악하고, 각 항목별로 세부 내용을 확인함으로써 데이터 리포지토리의 신뢰성을 확보하기 위한 내용을 분석하였다. 이를 기반으로 데이터 리포지토리의 신뢰성 향상과 관련하여 조직 인프라, 디지털 객체 관리, 기술 측면에서 논의를 진행하였다.

## 4. CTS 인증 요구사항 항목별 비교 분석 결과

본 장에서는 3장에서 분석 대상으로 선정한 DataON, PO.DAAC, Yareta, DARIAH-DE 4개의 리포지토리 신청서를 요구사항별로 비교분석한 내용을 기술한다. CTS의 인증 요구사항은 조직 인프라, 디지털 객체관리, 기술 인프라의 카테고리로 요구사항을 분류할 수 있으므

로, 각 카테고리별로 요구사항을 비교하면서 공통점과 리포지토리별 특징을 분석하였다(〈표 2〉 참조). 이를 통해 CTS 인증 신청 시 요구사항별로 리포지토리가 신뢰성 향상을 위해 갖추어야 할 정책, 시스템, 서비스 등에 대해서 파악하고자 한다.

### 4.1 조직 인프라

'R1 미션 및 범위'는 각 리포지토리의 조직의 사명과 임무에 대하여 작성하는 것으로 리포지토리를 운영하는 조직이 명확한 사명과 임무를 가지고 활동하는지를 확인한다. 해당 사명과 임무가 리포지토리에 링크되어 있는지도 확인한다. 모든 리포지토리와 이를 운영하는 기관에서는 연구데이터의 체계적인 관리를 통해 보유하고 있는 데이터에 대하여 장기적인 제공과 접근을 지원하고, 연구데이터를 관리할 수 있는 인프라를 제공한다. 한편, 각 리포지토리의 고유한 미션의 특징을 살펴보면 다음과 같다. DataON은 국가 R&D 지원 사업에서 산출되는 연구데이터를 대상으로 하며, PO.DAAC는 NASA의 지구과학 및 물리 해양학 데이터를 대상으로 한다. Yereta는 스위스 제네바 주의 연구원에서 산출되는 모든 연구데이터를, DARIAH-DE는 독일의 인문/사회 과학 분야의 연구데이터를 관리 대상으로 한다.

'R2 라이선스'는 사용 중인 라이선스, 사용 조건, 위반시 제재조치 정책이 잘 정의되어 있는지를 살펴본다. 모든 리포지토리에서 CC-BY 라이선스의 이용과 접근 권한 관리를 하고 있다. 각 리포지토리별 특징으로는 DataON에서는 연구데이터 이용 협약에 따라 이용 또는 제재조치를 시행한다. PO.DAAC는 소프트웨어 라이선스로 Apache 2.0을 준수한다. Yereta는 접근이나 이용 조건을 준수하지 않을 때는 LIPAD(Law on Public Information, Access to Documents, and Protection of Personal Data) 법에 따라 제재조치를 취한다. DARIAH-DE는 저작권 문제 발생 시, 삭제 조치한다.

'R3'은 접근의 지속성을 기술하며, 보존 기간 및 보존 책임과 이를 위한 계획 여부를 기술하도록 하고 있다. 모든 리포지토리에서 장기 보존 기간을 포함하는 정책과 장기 보존을 위한 절차가 수립되어 있다. 각 리포지토리별로 살펴보면, DataON은 출연연 법안으로 설립된 기관으로 영구적 운영을 보장한다. PO.DAAC는 모기관인 NASA의 기록 보존 일정과 데이터 활용도에 따른 첫번째 보존 기간 및 책임 수준에 대해 기술하고 있다.

'R4 기밀유지/윤리'는 공개위험이 있는 데이터에 대해서 어떤 절차와 활동이 필요한지를 기술하도록 하고 있다. 모든 리포지토리에서는 공개 위험이 있는 데이터에 대한 절차 및 활동과 관련한 규정을 수립하고 있다. 리포지토리별로 보면, DataON에서는 데이터 표준 계약 및 개인정보 보호 법안 등에 따라 공개 위험이 있는 데이터를 관리하고 있으며, PO.DAAC는 민감 데이터를 보관하지 않는다. Yareta의 경우 공개 위험이 있는 데이터는 등록 관련 모든 책임이 데이터 소유자에게 있다. DARIAH-DE의 경우 데이터 이용자가 모든 책임을 진다.

'R5 조직 인프라'는 기관 적합성 및 안정성과 관련한 내용을 기술한다. 모든 리포지토리와 운영기관은 전문 인력 및 기술적, 재정적 안

〈표 2〉 조직 인프라 부분의 요구사항 및 리포지토리별 구현 내용

| 항목 | 구분 | DataON | PO.DAAC | Yareta | DARIAH-DE |
|---|---|---|---|---|---|
| R1. 미션/범위 | Level | 4 | 4 | 4 | 4 |
| | 리포지토리별 특징 | 한국 국가 R&D 지원 사업 산출 연구데이터를 대상으로 함 | NASA의 지구과학 및 물리 해양학 데이터를 대상으로 함 | 스위스 제네바 주 연구원 산출 연구데이터를 대상으로 함 | 독일의 인문 사회 과학 분야 연구데이터를 대상으로 함 |
| | 공통점 | 연구데이터의 체계적 관리를 통한 장기적인 데이터 제공 및 접근 지원 / 연구 인프라 제공 | | | |
| R2. 라이선스 | Level | 4 | 4 | 4 | 4 |
| | 리포지토리별 특징 | 연구데이터 이용 협약을 기반으로 한 이용 및 제재 조치 시행 | Apache 2.0 준수 | 이용 조건 미준수시 LIPAD법에 따른 제재조치 시행 | 저작권 문제 발생 시 삭제 조치 시행 |
| | 공통점 | CC-BY 라이선스 이용 / 접근 권한 관리 | | | |
| R3. 접근의 지속성 | Level | 3 | 4 | 4 | 4 |
| | 리포지토리별 특징 | 출연연 법안으로 설립된 기관으로, 영구적 운영 | NASA의 기록 보존 일정 등 모기관의 데이터 활용도에 따른 운영 기간 결정 | Yareta 폐쇄 시 제네바 대학에 이관 | 데이터 재배치 또는 소유자 요청 사유 외 리포지토리에서 데이터 삭제 불가 |
| | 공통점 | 장기 보존을 위한 정책(보존 기간 포함)과 절차 수립 | | | |
| R4. 기밀/윤리 | Level | 4 | 4 | 4 | 4 |
| | 리포지토리별 특징 | 데이터 표준 계약 및 개인 정보 보호 법안에 따른 공개 위험 데이터 관리 | 민감 데이터는 보관하지 않음 | 공개 위험 데이터의 책임 권한을 데이터 소유주에 일임 | 데이터 이용의 모든 책임을 데이터 이용자에 일임 |
| | 공통점 | 공개 위험이 있는 데이터에 대한 절차 및 활동과 관련된 규정 수립 | | | |
| R5. 조직 인프라 | Level | 4 | 4 | 4 | 4 |
| | 리포지토리별 특징 | 한국 과기정통부 출연금으로 운영 / IT 서비스, 데이터 도메인 과학자 중심의 인력 구성 | NASA 제트 추진 연구소에서 운영 / 지구물리학, IT 전문가 중심의 인력구성 | 제네바 대학 IT 본부 e-Research 팀에서 운영 / CKAD IT 전문 자격증을 확보한 직원들로 구성 | NFDI HDC에서 운영 / 소프트웨어공학, 정보과학, 메타데이터설계, 디지털인문학, 데이터 마이닝 전문가 중심의 인력구성 |
| | 공통점 | 전문 인력 및 기술적·재정적 안정성을 기반으로 운영되며, 직원들의 전문 지식 확보 및 유지를 위한 교육을 제공함 | | | |
| R6. 전문가 지침 | Level | 4 | 4 | 4 | 4 |
| | 리포지토리별 특징 | 주기적인 협의체 운영을 통한 피드백 수렴 | 지구과학 분야 커뮤니티 참여 및 이용자 커뮤니티와 직접 소통 | 지역적/국제적 커뮤니티 참여를 통한 전문 지식 및 모범 사례 공유 | 연구데이터 인프라 컨소시엄 참여 |
| | 공통점 | 내부 전문가 활용을 통한 인프라 구축 및 기술적 협력 / 리포지토리 전문 분야 커뮤니티와의 소통을 통한 피드백 반영 및 협력 체계 구축 | | | |

정성을 기반으로 운영되며, 직원들의 전문 지식 확보를 위해 교육을 제공한다. 리포지토리별로 살펴보면, DataON은 국내 정부출연(연)으로 정부 자금으로 운영되며 도메인 과학, IT 서비스, 데이터 과학 등의 전문성을 갖춘 정규직원이 DataON을 운영하고 있다. PO.DAAC의 경우, NASA에서 운영하는 연구소로 지구물리학 IT 전문인력이 연간 12백만 달러의 예

산을 안정적으로 조달 받으며, 해당 직원들은 Earth Science Information Partners(ESIP), Apache 프로젝트 등 각 분야에서 선도할 수 있는 직원들이 참여하고 있다. Yareta는 Linux Foundation's Certified Kubernetes Application Developer(CKAD) 자격증 등을 확보한 IT 개발 전문가가 시스템을 개발하며, 도서관 직원이 시스템을 관리하는 역할을 나누어 수행하고, 2개의 대형 국가 프로젝트를 위한 비용을 지원받는다. DARIAH-DE는 독일의 국가 연구데이터 기반 시설(Nationale Forschungsdateninfrastruktur e. V.: NFDI) 산하에 있는 디지털 인문학센터(Humanities Data Centre: HDC)에서 운영하며, 운영자는 4명으로 작은 그룹에 속하지만, 각 직원들은 IT, 문헌정보, 데이터 마이닝, 디지털 인문학 등 다양한 전문 분야를 가지는 고경력전문가이다.

'R6 전문가 지침'은 리포지토리의 내부 자문가 또는 기술, 큐레이션, 데이터 과학 및 규율 전문가로 구성된 자문 위원회가 존재하는지, 전문가로부터 조언을 얻기 위한 리포지토리의 커뮤니케이션 방법은 어떠한지, 리포지토리가 피드백을 얻기 위해 지정 커뮤니티와 어떻게 소통하는지 등을 질문하고 있다. 본 연구에서 분석한 4개의 모든 리포지토리가 내부 전문가의 조언과 외부로부터의 조언을 받고 있었다. 4개 리포지토리 모두 내부 전문가의 조언을 통해 특히 인프라 구축 및 기술적(technical)인 측면에서 협력하는 것으로 파악되었고, 외부 전문가와의 커뮤니케이션 및 지정 커뮤니티와 소통하기 위한 방법은 리포지토리별로 상이하였다. DataON은 특히 지정 커뮤니티의 데이터 담당자들로 구성된 협의체를 주기적으로 운영함으로써 데이터 관리 및 리포지토리의 운영에 관한 피드백을 받으면서 소통하였으며, PO.DAAC는 'User Working Group'이라고 하는 이용자 커뮤니티를 조직하고 이용함으로써 이용자들의 피드백을 반영하고 있었다. Yareta와 DARIAH-DE는 외부 전문가 및 지정 커뮤니티와 소통하는 방법에 있어서 비슷한 양상을 보이고 있는데, 특정 분야의 다양한 커뮤니티 및 컨소시엄 등에 참여하고 협력을 진행하는 것으로 파악되었다.

### 4.2 디지털 객체 관리

R7은 데이터의 무결성과 진본성에 관하여 질문하고 있다. DataON, PO.DAAC, Yareta, DARIAH-DE는 모두 데이터 무결성과 진본성을 보장하기 위한 다양한 전략을 사용하고 있다. 이들 리포지토리는 공통적으로 데이터 무결성 확인을 위해 체크섬을 사용하며, 메타데이터와 데이터 변경 사항을 기록하고, 데이터 버전 관리 체계를 갖추고 있다. DataON은 DOIs와 함께 ORCID와 같은 식별자를 사용하고, Yareta는 파일 무결성을 위해 MD5, SHA1, SHA256 체크섬을 활용하며, PO.DAAC는 데이터 제공자가 생성한 체크섬을 사용해 구조적 무결성을 보장한다. DARIAH-DE는 EPIC2 Handle과 DataCite DOI를 통해 데이터에 고유 식별자를 부여한다. 차이점으로는 DataON과 PO.DAAC는 데이터 버전을 지원하고, Yareta와 DARIAH-DE는 현재 데이터 버전 관리를 지원하지 않으며, Yareta는 향후 지원 계획을 가지고 있다. 또한, PO.DAAC는 데이터 모델 구조와 일관성을 보장하기 위해 NASA CMR을 사용하며, DARIAH-DE는 데이터가 게시된 후 변경이 불가능하다는

〈표 3〉 조직 인프라 부분의 요구사항 및 리포지토리별 구현 내용

| 항목 | 구분 | DataON | PO.DAAC | Yareta | DARIAH-DE |
|---|---|---|---|---|---|
| R7. 데이터의 무결성과 진본성 | Level | 4 | 4 | 4 | 4 |
| | 리포지토리별 특징 | 데이터와 메타데이터의 모든 변경 사항 추적 및 데이터베이스 기록을 통한 버전 관리 | 메타데이터 규정 준수 검사기를 사용한 확인 및 데이터 속성으로 데이터 출처 추적과 모든 데이터 세트 활동의 포괄적 감사 추적 수행 | 데이터 및 메타데이터의 모든 변경 사항에 대한 감사 추적 사용 및 데이터에 대해 수행된 모든 작업 모니터링 / 기록 | 데이터 입수 시 메타데이터 입력 후 자동생성 메타데이터로 보완, 그러나 게시 후 데이터의 버전 관리는 지원하지 않음 |
| | 공통점 | 데이터의 무결성 및 신뢰성 유지를 위한 정책 수립 및 기능 구축 | | | |
| R8. 평가 | Level | 3 | 4 | 4 | 3 |
| | 리포지토리별 특징 | 성문화된 공식적인 수집 정책 부재 / 다양한 도메인의 데이터 속성을 반영하기 위한 공통 속성의 스키마 위주 | 선호하지 않는 형식의 파일인 경우 기탁을 불허함 / 관련 스키마에 대한 메타데이터 준수 여부를 관리자가 평가함 | DLCM tool을 활용하여 장기 보존 형식 준수 여부 및 메타데이터 유효성 검사 등 자동 수행 | Dublin Core 메타 가운데 3개 필드만 필수(제목, 작성자, 라이선스) |
| | 공통점 | 메타데이터 스키마의 준수 여부를 자동 및 수동으로 확인 및 선호 파일 형식 사용 | | | |
| R9. 문서화된 저장 절차 | Level | 4 | 4 | 4 | 4 |
| | 리포지토리별 특징 | 리포지토리 내부 구성 소프트웨어 및 백업 프로세스에 대한 자세한 문서 관리 / 서버 디스크 및 백업 시스템 상태 감독을 포함한 포괄적인 모니터링 시스템 이용 | 보안을 위한 클라우드 솔루션을 통합, 지속적으로 작동하는 모니터링 기계실 운영 / 시스템 모니터링 및 정기적 업데이트를 통한 지속적 데이터 가용성 보장 | 아카이브 저장과 관련된 모든 절차의 문서화 및 저장 위치를 서술하는 특정 보안 정책 구축 | 동적 및 정적 데이터 세트를 위한 별도 가상 연구 환경에서 자동 가능 / 손실 데이터 복구 프로토콜을 포함한 포괄적 위험 관리 계획 실행 |
| | 공통점 | 데이터 관리를 위한 문서화 시행 및 전략 수립 / 데이터 소실 방지를 위한 분산 저장 | | | |
| R10. 보존 계획 | Level | 3 | 4 | 3 | 3 |
| | 리포지토리별 특징 | 마이그레이션과 정규화 전략 통합 | NASA가 정한 높은 수준의 요구 사항 준수 및 보존 책임 이행 | OAIS 규정 준수 및 관리 책임을 포함한 디지털 저장소의 모든 속성을 다루는 보존 정책 보유 | 괴팅겐 대학의 개방형 접근 전략에 부합하는 보존 정책 보유 |
| | 공통점 | 문서화된 보존 전략 수립 및 보존 책임 명시 / 기술적 노후화 해결을 위한 사전 전략 수립 | | | |
| R11. 데이터 품질 | Level | 4 | 4 | 4 | 3 |
| | 리포지토리별 특징 | 무결성을 기본 원칙으로 삼음 / 연구데이터 생산자의 학문적 지식 및 연계 기관의 평가 신뢰 | NASA wide Unified Metadata Models를 이용 및 해당 분야 스키마 사용 | 유효한 스키마가 입력된 경우에만 데이터 기탁 가능 / 연구자가 별점을 제공하는 형식으로 피드백 제공 | 최소한의 메타데이터 값을 필수로 함 (3개) / 커뮤니티 표준 및 기술적인 변화 반영을 통한 메타데이터 선택 입력 값 조정 |
| | 공통점 | 메타데이터 중심의 품질 관리 수행 | | | |
| R12. 워크 플로우 | Level | 4 | 4 | 4 | 4 |
| | 리포지토리별 특징 | 직접 기탁 데이터와 연계하여 수집한 데이터로 나누어 관리 | 식별, 준비, 통합, 운영 단계에 따라 관리함 | 표준과 도구가 명시된 관련 정책 수립 | 데이터 관리에 필요한 모든 과정 문서화 / 데이터 처리에 관한 기탁자 및 이용자와의 의사 소통 방법은 언급하지 않음 |
| | 공통점 | 각 기관에서 이용 가능한 가이드라인 또는 체크리스트에 따라 워크플로우 수행 | | | |

| 항목 | 구분 | DataON | PO.DAAC | Yareta | DARIAH-DE |
|---|---|---|---|---|---|
| R13. 데이터 발견 및 식별 | Level | 4 | 4 | 4 | 4 |
| | 리포지토리별 특징 | DataCite, DCAT, Schema.org를 참조한 자체 스키마 활용 | 지구과학 분야 표준 EOSDIS 메타데이터 활용 | DataCite 메타데이터 활용 | Dubline Core 메타데이터 활용 |
| | 공통점 | '검색 기능', 'Facet 검색', '컬렉션(제공처 검색)' 제공 / DOI를 통한 데이터셋 게시 | | | |
| R14. 데이터 재이용 | Level | 4 | 4 | 4 | 3 |
| | 리포지토리별 특징 | 공통 메타데이터 이외의 정보는 분야별 '특성 정보'로 관리 / 데이터 재이용을 위한 데이터 파일 설명 자료 제공 | 데이터셋 뿐만 아니라 사용한 알고리즘, 소프트웨어 등 필요한 도구 및 부가정보 함께 제공 | 공통 메타데이터 이외에 이용자 메타데이터 연결 기능 제공 | 디지털 신뢰성을 위해 Nestor 기준 이용 / 포맷 변화 반영을 위해 커뮤니티 표준 모델 개발 및 모니터링 수행 |
| | 공통점 | 지속적인 모니터링과 재평가를 통해 시간에 따른 포맷 변화 지원 체계를 지원함 | | | |

점에서 특이하다. 이들 리포지토리는 각각의 운영 환경과 데이터 특성에 맞는 최적의 방법을 적용하고 있다.

'R8 문서화된 저장 절차'의 경우, 평가 절차에 대한 정책 및 기능의 여부를 확인하며, 컬렉션 정책, 메타데이터 제공 및 준수여부, 선호 파일 운영, 데이터 제거 프로세스와 같은 항목을 검사한다. DataON과 DARIAH-DE는 준수 레벨 3을 획득하였으며, PO.DAAC와 Yareta는 준수 레벨 4를 획득하였는데, DataON의 경우 성문화된 공식적인 수집 정책이 존재하지 않고, 메타데이터 스키마의 발견성에 초점을 맞춘 매우 일반적 스키마라는 문제점이 있으며, 다양한 도메인의 데이터 속성을 반영하기 위한 공통 속성의 스키마 위주로 작성되어 평가의 한계가 있다. DARIAH-DE의 경우는 컬렉션 개발 정책은 있으나, Dublin Core 메타데이터 스키마를 사용하고 있으며, 3개 필수 필드만을 지정함으로써 장기적으로 다양한 지정 커뮤니티에서 적절하고 이해 가능한 수준을 보장하기에는 다소 부족한 것이 사실이다. 메타데이터의 준수여부에 대해서는 모든 리포지토리가 자동 또는 수동으로 준수여부를 확인하고 있으며, 조직의 미션이나 컬렉션 프로파일에 속하지 않는 디지털 객체에 대해 PO.DAAC는 기탁을 불허하는데, 이는 NASA 지원을 받는 과학팀에서 임무지향형 데이터셋을 생성하기 때문으로 판단된다. 그리고, 선호파일형식(Preferred File Format)의 경우, 각 리포지토리가 다양한 명칭으로 선호파일형식을 정의하고 관리하고 있는데, 특히 DARIAH-DE의 경우 특정파일형식을 권장하지만 선호 형식 제공 시 TEI 또는 XML과 같은 보충서비스를 제공하고 있고, PO.DAAC의 경우 선호파일형식이 아닌 경우 기탁을 불허하고 있다. 그리고 모든 리포지토리가 컬렉션에서 아이템을 제거할 때 영구식별자와 메타데이터는 유지하고 있다.

R9은 저장 절차를 문서화하고 있는 지에 대해 질문하고 있다. DataON, PO.DAAC, Yareta, DARIAH-DE는 모두 CTS 요구 사항에 따라 문서화된 저장 절차를 철저히 수행하지만, 접근 방식에는 차이가 있다. 공통적으로 이들 리포지토리는 백업 시스템을 운영하고, 데이터 손실을 방지하기 위한 일관성 확인과 리스크 관

리 기법을 적용하며, 재해 복구 계획을 마련하고 있다. DataON과 PO.DAAC는 서비스 및 백업 저장소를 구분하고 정기적인 백업과 복구 훈련을 수행한다. PO.DAAC는 AWS 클라우드를 활용해 데이터 중복과 백업을 관리하며, Yareta는 지리적으로 분산된 두 개의 저장소에 데이터를 보관하여 재난에 대비한다. DARIAH-DE는 동적 저장소와 정적 저장소로 나누어 데이터를 관리하며, RAID6와 같은 기술을 사용해 저장 매체의 열화를 방지한다. 이들 리포지토리는 각기 다른 운영 환경과 데이터 특성에 맞춰 최적의 저장 절차를 문서화하고 관리하고 있다.

R10에서는 보존 계획에 대해 질문하고 있다. DataON, PO.DAAC, Yareta, DARIAH-DE는 모두 장기 보존 계획을 문서화하고 실행 중이지만, 접근 방식에는 차이가 있다. 공통적으로 이들 리포지토리는 데이터의 장기적 접근성과 사용성을 보장하기 위해 백업 시스템과 데이터 형식 변환 절차를 포함한 보존 전략을 채택하고 있다. DataON은 데이터와 메타데이터를 안전한 미디어에 저장하고, 데이터 포맷 변환을 통해 장기 보존을 보장하며, 위험 관리 단계를 체계적으로 진행한다. PO.DAAC는 NASA의 보존 정책에 따라 데이터 형식 변환과 클라우드 저장소를 활용하여 데이터 보존을 관리한다. Yareta는 OAIS 원칙을 따르며, 비트스트림 보존을 통해 데이터의 장기 접근성을 유지하지만, 독점 파일 형식의 변환에 대한 도전을 인식하고 있다. DARIAH-DE는 개방형 파일 형식의 장기 보존을 보장하고, 법적 및 규제 프레임워크를 준수하여 데이터 접근성을 유지한다. 각 리포지토리는 자체 운영 환경과 데이터 특성에 맞는 최적의 보존 전략을 적용하고 있다.

'R11 데이터 품질'에서는 리포지토리가 데이터 및 메타데이터 품질에 어떻게 접근하고 있는지, 리포지토리에 기탁된 데이터의 완전성과 이해가능성을 보장하기 위해 품질 관리를 어떻게 진행하고 있는지, 지정 커뮤니티가 데이터와 메타데이터에 대한 의견을 제시하거나 평가할 수 있는지 등을 기술한다. DARIAH-DE를 제외한 3개의 리포지토리가 최고점인 4점을 받았으며, DARIAH-DE는 아직 구축이 진행 중이어서 레벨 3을 받았다. 전체적으로 4개의 리포지토리 모두 데이터 품질과 관련해서 메타데이터를 중심으로 진행되는 것으로 나타났다. 데이터 품질 부분에서는 DataON과 DARIAH-DE가 유사하며, Yareta와 PO.DAAC가 유사하다고 볼 수 있다. 이는 Yareta와 PO.DAAC는 특정 연구 분야의 리포지토리이지만, DataON과 DARIAH-DE는 각각 과학기술분야와 인문학 전반을 다루고 있는 광범위한 주제의 리포지토리인 것으로 파악된다. 먼저 DataON은 데이터 품질 정책은 없지만 '무결성'을 가장 중요한 원칙으로 삼아 기탁된 데이터 및 연계된 데이터가 전달하려는 내용을 최대한 왜곡 없이 제공할 수 있도록 노력한다. 또한, 연구데이터라는 특성에 기반하여 기탁된 데이터 자체의 품질에 대한 평가를 진행하지 않고 있으며, 이는 데이터를 생산한 연구자의 학문적 지식과 연계를 진행한 기관의 데이터 담당자의 평가를 신뢰하여 데이터 품질을 기대한다. 데이터 품질은 메타데이터에 집중하여 유효성 등을 평가하고 있으며, 대신에 협의체 및 커뮤니티 모니터링 등을 통해 데이터 자체에 대한 품질을 관리할 수 있도록 노력하고 있다. DARIAH-DE의 경우 인문학 전반의 데이터를 다루고 있기 때문에 최소한의

메타데이터 값을 필수로 가지고 있다. 현재 더블린코어 심플 메타데이터를 기본으로 하여 3개의 필수 스키마를 가지고 있는데, 최소한의 필수 메타데이터 값으로는 데이터의 완전성 및 이해가능성을 보장하기 어려운 한계가 존재하기 때문에 DARIAH-DE는 커뮤니티 표준 및 기술적인 변화를 지속적을 모니터하여 반영할 수 있게 함으로써 선택 입력 메타데이터 스키마 값을 조정한다. 그러나 버전 관리를 제공하지 않고, 관련 저작물에 대한 인용 및 인용 색인 링크 등을 제공하지 않아 해당 연구데이터 품질 확보에 한계를 가지고 있다. Yareta와 PO.DAAC는 모두 특정 분야에 집중한 리포지토리이기 때문에 기본 메타데이터뿐만 아니라 해당 분야에서 필요로 하는 메타데이터 값을 정리하여 추가로 입력할 수 있도록 독려하고 있다. 특히, 수집 및 관리하는 연구데이터의 분야가 확실하기 때문에 대상으로 삼는 지정 커뮤니티 역시 해당 분야의 전문 커뮤니티이며, 리포지토리 담당자는 해당 분야의 커뮤니티에 적극적으로 참여하여 피드백을 받고 이를 리포지토리 운영에 반영하고 있다.

R12에서는 워크플로우와 비즈니스 프로세스를 설명해야 한다. 4개의 리포지토리 모두 각 기관에서 이용 가능한 가이드라인 또는 체크리스트에 따라 워크플로우를 수행할 수 있는데, 워크플로우는 각 기관별 리포지토리의 관리 흐름을 설명하기 때문에 명확한 기준에 따라 요구사항을 만족하는지에 대해서 파악이 어려운 부분이 있다. DataON과 PO.DAAC는 각 단계별로 제공되는 가이드라인에 따라 워크플로우가 진행되며, Yareta는 표준 및 도구까지 명시하는 가이드라인을 제공하고 있다. 그리고 DARIAH-DE 역시 데이터 관리에 필요한 모든 과정을 명시한 문서를 기준으로 데이터를 관리하고 있다.

'R13 데이터의 발견과 식별'에서는 4개의 리포지토리 모두 '검색 기능', 'Facet 검색', '컬렉션(제공처) 검색' 등을 제공하고, 영구식별자로 DOI를 통해서 데이터셋을 게시한다. 각 리포지토리별로 살펴보면, 메타데이터의 경우, DataON은 DCAT, DataCite, Schema.org 등을 참조한 자체 메타데이터를, PO.DAAC는 지구과학 분야 표준인 EOSDIS metadata를 따르고 있으며, Yareta는 DataCite를 따른다. DARIA-DE의 경우, R13에 언급하진 않았지만, R14에는 더블린 코어에서 정의한 메타데이터로 기술한다고 되어 있다.

'R14 데이터 재이용'에서는 모든 리포지토리에서 지속적인 모니터링 및 재평가를 통해 시간에 따른 포맷 변화 지원 체계를 지원하고 있다. 리포지토리별 특징을 살펴보면, 지속적인 이해 가능성을 위해 DataON은 메타데이터 이외의 연계 사이트에서 제공하는 메타데이터가 존재 시, 특성 정보로 관리하고 있으며, 이해를 돕기 위해 데이터 파일 설명 자료도 추가할 수 있다. PO.DAAC는 데이터셋뿐만 아니라 사용한 알고리즘, 소프트웨어 등 필요한 도구들과 부가 정보도 함께 제공한다. Yareta는 데이터 이해를 돕기 위한 메타데이터를 정의하여 관리하며, DARIA-DE의 경우 신뢰성을 위해 Nestor 제공 기준을 따른다. 지정된 커뮤니티의 포맷 사용을 지원하기 위해 DataON에서는 XML 형태로 메타데이터를 저장하고, 시공간정보는 OpenStreetMap등을 통해 가시화하여 보여줄 수 있도록 well-known text로 저장한다. PO.DAAC에서는 NASA Earth science가 승인 및 표준

화한 데이터 포맷 사용을 권장하며, Yareta에서는 공통 메타데이터 이외에 이용자 정의 메타데이터 또한 연결해서 이용할 수 있도록 지원한다. DARIA-DE의 경우, Text+ 프로젝트의 요구사항을 준용하며, Geo Browser를 통해 시공간 정보는 가시화한다.

### 4.3 기술

R15의 경우, 모든 리포지토리에서 국제적 혹은 커뮤니티 등의 다양한 내/외부 표준 및 규정을 준수하고 있으며, 지정 커뮤니티 및 이용자의 요구 사항 충족을 위해 충분한 대역폭을 보장하기 위해 높은 가용성 확보를 위해 노력하고 있으며, 다양한 형태로 재난 또는 재해로부터 서비스를 신속하게 복구하기 위하여 백업 및 복구 체계와 원격지 재해복구 체계를 구축하고 운영 중이다. 인프라 개발 계획의 경우 DataON은 매년 예산 투입을 통해 추진하고 있으나, 다른 리포지토리의 경우는 수요에 기반하여 추진하고 있다.

R16의 경우, 모든 리포지토리가 IT 보안 시스템과 보안 관련 역할을 수행하는 직원을 보유하고, DARIAH-DE의 경우 인프라 및 운영에 대한 시나리오 기반의 내부 위험관리 계획을 수립 및 운영하고 있다. 인프라 등의 시설물에 대한 보안의 경우, DataON, Yareta 및 DARIAH-DE는 각 리포지토리 관리 기관에서 보안 조치를 수행하나, PO.DAAC의 경우는 JPL(제트 추진 연구소)에서 클라우드 도구 및 서비스 보호 책임을 지고 있으며, Amazon 웹 서비스 보안과 협력하고 있다. 외부로부터의 접근에 대해서는 모든 리포지토리가 방화벽 등을 통하여 액세스를 제어하고 있으며, 특히 이용자 인증에 있어서 DataON의 경우 패스워드 방식과 협력 기관은 연합인증방식 두 가지를 사용하고 있다.

〈표 4〉 조직 인프라 부분의 요구사항 및 리포지토리별 구현 내용

| 항목 | 구분 | DataON | PO.DAAC | Yareta | DARIAH-DE |
|---|---|---|---|---|---|
| R15. 기술 인프라 | Level | 4 | 4 | 4 | 4 |
| | 리포지토리별 특징 | 인프라 이중화 / 정보보호, 보안 가이드라인 준수, 웹취약점 점검 | OGC, NASA UMM 등과 같은 커뮤니티 표준 및 규칙 준수 | OAIS, FAIR 원칙, OAI-PMH, DataCite 메타데이터, METS, PREMIS 등의 표준 준수 | Shibboleth를 통한 인증 기능 구현 |
| | 공통점 | 국제적 혹은 커뮤니티 등 다양한 내·외부 표준 및 규정 준수 / 소프트웨어 및 인프라 유지관리 / 재해복구 체계 구축 운영 | | | |
| R16. 보안 | Level | 4 | 4 | 4 | 4 |
| | 리포지토리별 특징 | 패스워드 방식 및 연합인증 방식 두 가지의 이용자 인증 방식 활용 | JPL의 시스템 보안 및 NGAP의 클라우드 도구 및 서비스 보호 이용 / Amazon Web Service West 클라우드 인프라 보호지침을 따름 | 제네바 대학교의 '정보시스템에 대한 보안 정책'에 따라 보안 수행 / 데이터 보관을 위해 조직 단위의 SWITCHID를 Yareta에 연결함 | SUB 및 GWDG에서 전체 시스템 모니터링 및 조치 / 데이터 접근을 위한 다단계 인증 활용 |
| | 공통점 | IT 보안 시스템 및 보안 관련 담당 직원 보유 및 위험분석 활동 / 요구 보안 수준 운영 및 안전한 액세스를 위한 인증 및 승인 | | | |

이와는 달리 DARIAH-DE는 다단계 인증을 적용하고 있다.

## 5. 데이터 리포지토리의 신뢰성 향상을 위한 방안 제언

5장에서는 CTS 인증을 획득한 총 4개의 리포지토리를 선정하여 CTS 인증 요구사항 항목별로 해당 리포지토리가 어떻게 구현되어 있는지 비교 분석한 4장의 결과를 가지고, 다음과 같이 데이터 리포지토리의 신뢰성 향상을 위한 방안을 제안한다. 이를 위해 CTS 신청서의 대분류(조직 인프라, 디지털 객체 관리, 기술)로 구분하여 다음과 같은 방안을 제시한다.

### 5.1 조직 인프라 구축 방안

CTS에서는 신뢰할 수 있는 데이터 리포지토리가 '조직 인프라' 측면에서 갖추어야 할 사항에 대해 CTS 신청서의 R1부터 R6(미션/범위, 라이선스, 접근의 지속성, 기밀/윤리, 조직 인프라, 전문가 지침)에서 다루고 있다. 4장에서 비교분석한 데이터 리포지토리의 각 사항에 대한 답변을 분석한 결과를 가지고 다음과 방안을 제시한다.

데이터 리포지토리의 신뢰성을 향상시키기 위해서는 해당 리포지토리를 운영하는 조직 인프라를 체계적으로 구축하는 것이 선행되어야 한다. CTS는 리포지토리를 운영하는 기관의 안정성과 영속성, 그리고 분야에 대한 적합성을 평가하고 전문 IT 서비스를 지원할 수 있는지 여부와 이들에 대한 교육의 진행 여부 등을 파악하여 해당 리포지토리가 체계적으로 운영되고 있는지를 평가한다. 이를 기반으로 리포지토리 운영 기관은 해당 리포지토리의 분야 및 특성에 맞는 조직적 구성을 갖춤과 동시에 이를 지속적이고 안정적으로 운영할 수 있는 재정적, 구성적, 기술적 기반을 갖추어야 한다.

데이터 리포지토리를 운영하는 기관은 데이터의 장기 보존을 위한 조직 체계를 구성해야 한다. 조직 인프라 구축은 데이터 리포지토리를 운영하는 임무와 역할을 명확하게 하는 것이기 때문에 궁극적으로 재정적 안정성을 기반으로 전문인력을 구성하고 내외부 협력 방안을 마련해야 한다. 지구과학, 물리, 해양 등 전문 분야의 데이터 리포지토리는 연구자 커뮤니티의 필요에 의해 지속성을 유지할 수 있지만, DataON과 같이 다양한 연구분야를 포괄하는 통합 데이터 리포지토리의 경우에는 운영 기관의 사명과 조직의 업무 적합성도 함께 갖추어야 지속가능성을 확보할 수 있다.

조직 인프라는 내부의 조직 구성만을 의미하는 것이 아니라 외부와의 소통을 포함하는데, 특히나 데이터 리포지토리의 신뢰성을 확보하기 위해서는 해당 리포지토리의 데이터를 제공 및 이용하는 커뮤니티와의 의사소통이 중요하다. 본 연구에서는 다양한 외부 커뮤니티와의 소통 방법을 확인할 수 있었는데, 크게는 이용자 커뮤니티와의 소통, 특정 분야 전문가 커뮤니티와의 소통, 데이터 관리 담당자로 구성된 협의체를 운영하여 소통하는 것으로 나누어서 이해할 수 있었다. 이용자 커뮤니티는 해당 리포지토리를 이용하는 이용자로 구성된 집단으로써, 리포지토리는 이용자 커뮤니티로부터 데이터 및 리포지토리 이용에 대한 의견을 수집

할 수 있고 이를 데이터 품질관리 및 이용자 편의 제고를 위한 방안에 반영할 수 있다. 특정 분야 전문가 커뮤니티는 일반적으로 데이터 리포지토리가 해양생물이나 천문학 등 특정 분야에 관련된 경우일 때 특히나 많이 이용되는 의사소통 방법인데, 전문가 커뮤니티는 데이터 이용자이면서 동시에 데이터 생산자로서 다양한 의견을 제시할 수 있고, 데이터 생산자인 연구자로서 데이터 품질관리에 대한 도움을 제공할 수 있다. 마지막으로 데이터 관리 담당자로 구성된 협의체는 특히나 DataON에서 이용하는 의사소통 방법인데, 이는 DataON이 국가연구데이터플랫폼으로서 다양한 출연(연)의 데이터를 연계하는 과정에서 그들의 요구 및 어려움 등의 의견을 반영하기 위한 방법으로 이용하고 있다. 이처럼 데이터 리포지토리를 운영하는 기관은 제도 마련, 전문 인력 양성, 이용자의 요구사항 반영, 디지털 객체 관리 방안, 기술 발전 등을 이유로 신뢰성 향상을 위한 협의체나 위원회를 구성할 수 있고, 이는 데이터를 직접 관리하는 실무자들과의 소통과 협력을 가능하게 하여 결과적으로 리포지토리의 운영에 도움이 될 수 있다.

### 5.2 디지털 객체 관리 방안

CTS에서는 신뢰할 수 있는 데이터 리포지토리가 '디지털 객체 관리' 측면에서 갖추어야 할 사항에 대해 CTS 신청서의 R7부터 R14(데이터의 무결성과 진본성, 평가, 문서화된 저장절차, 보존 계획, 데이터 품질, 워크플로우, 데이터 발견 및 식별, 데이터 재인용)에서 다루고 있다. 4장에서 비교분석한 데이터 리포지토리의 각 사항에 대한 답변을 분석한 결과를 가지고 다음과 방안을 제시한다.

데이터의 수집에서 재이용까지 진행되는 데이터 수명주기의 여러 단계에서 안정성 검사를 철저하게 실시해야 한다. 각 단계를 거치면서 디지털 객체가 변경되거나 손상되지 않았는지 확인해야 하며, 데이터 리포지토리는 데이터와의 모든 상호 작용을 기록으로 남김으로써 포괄적이고 상세한 데이터 변경 추적을 유지해야 한다. 이를 기반으로 이용자와 관리자 모두 데이터의 기록을 확인할 수 있도록 하고 신뢰성을 확보할 수 있다.

데이터와 메타데이터 관리를 진행함에 있어서 리포지토리는 모든 변경 사항을 명확하게 문서화하여 버전을 관리해야 하며, 필요한 경우 이전 버전으로 되돌릴 수 있는 기능을 제공해야 한다. 데이터 저장 및 메타데이터 관리를 진행함에 있어서는 가능한 한 국제 표준을 채택해야 하는데, 광범위하게 이용되는 국제 표준을 채택 및 사용함으로써 신뢰성을 확보할 수 있다. 메타데이터의 경우, 다루는 데이터의 분야가 광범위할 경우에 특정한 하나의 표준으로는 다양한 데이터 특성을 반영하기 어려운 경우가 존재하는데, 이러한 경우에는 기본적으로는 국제적 표준을 최대한 반영하면서, 특정 분야 데이터에 대한 이해를 도울 수 있는 특징을 갖는 메타데이터에 대한 입력을 추가적으로 확보함으로써 데이터에 대한 이해도와 접근성을 높일 수 있도록 해야 한다.

리포지토리는 데이터 수집을 위해 성문화된 수집 정책을 가지고, 기탁된 메타데이터와 데이터에 대하여 완전성, 이해가능성, 장기 보존을 위한 요건을 충족하는지 검토해야 한다. 메타데

이터는 지정 커뮤니티 또는 이용자가 데이터를 이해하기 위해 충분한 내용을 갖추어야 하며, 리포지토리에서 필수적으로 요구되는 보존 메타 속성을 정의하여 포함해야 한다. 뿐만 아니라, 데이터를 장기 보존하여 이용자가 데이터를 지속적으로 이용할 수 있도록 파일의 속성을 관리하고 필요한 경우 마이그레이션을 수행해야 한다. 그리고 데이터를 다루는 전 주기에서 자동 또는 수동으로 데이터와 데이터 처리 과정을 모니터링하고 관리할 수 있는 체계를 구축해야 한다.

추가적으로 리포지토리의 신뢰도를 높이기 위해서는 엄격한 보존 정책을 구축하고 실행해야 한다. 리포지토리는 데이터 백업, 마이그레이션 및 복구 절차를 설명하는 포괄적이고도 명확한 문서화된 보존 계획을 수립해야 하며, 이를 기반으로 기술적 변화 및 외부의 변화가 발생하더라도 지속적으로 데이터에 접근하고 이용할 수 있도록 보장해야 한다. 리포지토리는 물리적 재해 또는 시스템 오류 등으로 인한 데이터 손실을 방지하기 위해 물리적으로 분산된 위치에 여러 데이터 복사본을 저장하여야 하며, 향후 데이터 포맷과 이를 지원하는 소프트웨어 및 운영체제 등 전반적인 인프라 변화에 대응하기 위한 정책 또는 관리 방안 역시 구축되어야 한다.

### 5.3 기술 인프라 방안

CTS에서는 신뢰할 수 있는 데이터 리포지토리가 '기술' 측면에서 갖추어야 할 사항에 대해 CTS 신청서의 R15와 R16(기술 인프라, 보안)에서 다루고 있다. 4장에서 비교분석한 데이터 리포지토리의 각 사항에 대한 답변을 분석한 결과를 가지고 다음과 방안을 제시한다.

데이터 리포지토리의 소프트웨어는 다양한 소프트웨어의 지속적인 유지 및 관리를 위해 문서화할 필요가 있으며, 안전한 저장소에 소프트웨어를 저장하여 인벤토리 및 버전 등을 지속적으로 관리해야 한다. 시설 및 리포지토리를 구성하는 시스템 등과 같은 인프라 역시 정전, 화재 등의 재난 또는 재해로부터 가용성을 유지하고, 장기 보존에 적합하도록 일련의 문서화된 정책 및 절차가 필요하며, 주기적으로 이에 대한 점검을 실시해야 한다. 이와 같은 과정에서 관련된 기술 표준 또는 규정을 참조하거나 준수해야 한다. 또한, 지정 커뮤니티 및 이용자 요구사항의 변화, 그리고 기술적인 변화에 대한 지속적인 모니터링과 이를 리포지토리에 반영할 수 있는 체계를 구축해야 한다.

리포지토리는 잠재적인 내·외부의 위협을 분석하고, 위험에 대한 평가를 진행함으로써 일관된 보안 시스템을 구축해야 한다. 다양한 위협 시나리오(악의적 행동, 인적 오류 또는 기술적 장애)를 설정하고, 시나리오에 대한 장애 영향 범위 및 대응 조치를 문서화함으로써 이를 주기적으로 점검해야 한다. 또한, 리포지토리는 보안 관련 표준을 참조하거나 준수하여 보안 체계를 구성하고 보안 관련 활동을 지속적으로 수행해야 한다.

## 6. 결 론

데이터의 가치와 데이터 리포지토리의 역할이 점점 중요해지면서, 데이터 리포지토리의 신

뢰성도 주목받고 있다. 데이터 리포지토리의 신뢰성은 보유하고 있는 데이터의 무결성을 보장하고, 데이터의 기탁과 이용 결정에 영향을 미치며, 데이터 리포지토리와 이해관계자 간의 신뢰를 구축하고 유지하는데 중요하다. CTS 인증은 국제적으로 데이터 리포지토리의 신뢰성을 인증하며, 데이터 관리 및 보존에 있어서 표준 준수를 확인하는 중요한 도구이다.

본 연구는 데이터 리포지토리의 신뢰성 향상을 위한 방안을 제안하고자 CTS 인증을 받은 KISTI의 DataON을 포함하여 총 4개의 국내외 데이터 리포지토리를 사례로 선정하여 CTS에서 명시하고 있는 리포지토리의 신뢰성 확보를 위해 요구되는 필수 요건들을 기준으로 삼아 각 리포지토리의 구현 내용을 비교 분석하였다. 본 연구에서 서술한 'CoreTrustSeal Trustworthy Data Repositories Requirements' 내용과 4개의 리포지토리 신청서를 요구 필수 사항별로 비교 분석한 내용은 CTS 인증을 신청하려는 데이터 리포지토리들이 참고할 수 있을 것이다.

CTS 인증을 받은 4개의 데이터 리포지토리 신청서를 요구 필수 사항별로 비교 분석해 본 결과, 모든 리포지토리는 각 사항별로 명확하게 답변하고, 그 답과 관련된 링크를 그 근거로 제시하고 있었다. 또한 4개의 신청서는 대부분의 사항에서 4점을 받았으나, 3점을 받은 사항도 있었다. CTS 신청서에서 최고점인 4점을 받은 답변 내용을 바탕으로 하여 5장에서 제안한 데이터 리포지토리의 신뢰성 향상을 위한 방안을 요약해 보면 다음과 같다.

첫째, 데이터 리포지토리의 신뢰성을 향상시키기 위해서는 체계적인 조직 인프라 구축이 우선되어야 한다. 이를 위해서 기관의 안정성, 영속성, 분야 적합성, 전문 IT 서비스 지원 여부 등을 평가하고, 재정적, 구성적, 기술적 기반을 마련해야 한다. 또한, 데이터 리포지토리의 신뢰성을 확보하기 위해서는 이해관계자들과의 의사소통이 중요하다. 데이터 생산자, 데이터 이용자, 데이터 분야의 전문가, IT 기술자 등 다양한 이해관계자들과 지속적인 소통과 협력을 해야 한다. 이를 통해 그들의 의견을 수렴하고 반영하는 협의체를 운영하여 그들의 필요를 파악하고, 이용자의 만족도와 데이터 품질을 향상시키는 등 궁극적으로 데이터 리포지토리와 이해관계자들 사이의 신뢰를 향상시킬 수 있다.

둘째, 디지털 객체 관리 측면에서는 데이터의 수명주기 전반에 걸쳐 안정성 검사를 수행하고, 데이터를 변경하거나 처리한 모든 활동을 기록으로 남겨야 한다. 예를 들어, 데이터와 메타데이터 관리를 명확하게 문서화하고 국제 표준을 채택하여 신뢰성을 확보해야 한다. 또한, 데이터 리포지토리는 기관의 수집 정책과 보존 정책을 수립하고 그러한 정책에 따라 데이터의 수집과 보존 활동을 수행해야 한다.

셋째, 기술적 측면에서는 소프트웨어를 지속적으로 유지관리하고 그러한 활동에 대해 기록으로 남겨야 한다. 또한 소프트웨어는 안전한 저장소에 저장하고, 인벤토리와 버전을 관리해야 한다. 더 나아가 인프라의 가용성을 유지하고, 기술적인 표준과 규정을 준수해야 한다. 보안을 위하여 보안 시스템을 구축하고 다양한 위협 시나리오에 대응할 수 있는 위기 관리 계획을 준비하고 필요시 실행한다.

본 연구는 CTS 인증 신청서를 비교분석하여 데이터 리포지토리의 신뢰성 향상을 위한

방안을 제안한 연구로서 데이터 리포지토리의 신뢰성이 점점 중요해지는 상황에서, 향후 이와 관련된 주제의 후속 연구들이 나오기를 기대한다. 예를 들어, 인증이 신뢰에 미치는 영향에 대한 연구가 수행될 수 있는데, CTS와 같은 인증이 데이터 리포지토리의 신뢰성과 평판에 미치는 장기적인 영향을 조사할 수도 있고, 신뢰도 향상에 있어서 다양한 인증 표준의 효과를 비교해보는 연구도 가능하다. 다른 한편으로, 데이터 리포지토리의 이용자 관점에서, 다양한 이해관계자 그룹(예: 연구원, 데이터 관리자, 정책 입안자)이 데이터 리포지토리에 대한 신뢰를 어떻게 인식하는지 조사할 수 있다. 더 나아가, 다양한 이해관계자 간의 신뢰 수준에 영향을 미치는 요인을 분석하는 연구를 기대해 볼 수 있다.

## 참 고 문 헌

김주섭, 양성준, 김선태 (2022). 데이터 리포지토리 인증 체계 분석 및 인증 전략에 관한 연구: Coretrustseal을 중심으로. 한국문헌정보학회지, 56(2), 209-229. https://doi.org/10.4275/KSLIS.2022.56.2.209

Azeroual, O. & Schöpfel, J. (2021). Trustworthy or not? Research data on COVID-19 in data repositories. Libraries. Digital Information, and COVID. 2021, 169-182. https://doi.org/10.1016/B978-0-323-88493-8.00027-6

Broekstra, R. (2021). Values of Trust and Participation in Scientific Data Repositories. University of Groningen.

CoreTrustSeal (2024). About. Available: https://www.coretrustseal.org/about/

Corrado, E. M. (2019). Repositories, trust, and the CoreTrustSeal. Technical Services Quarterly, 36(1), 61-72. https://doi.org/10.1080/07317131.2018.1532055

Donaldson, D. R. (2020). Certification information on trustworthy digital repository websites: A content analysis. PLoS ONE, 15(12), e0242525. https://doi.org/10.1371/journal.pone.0242525

Downs, R. R. (2021). Improving opportunities for new value of open data: Assessing and certifying research data repositories. Data Science Journal, 20, 1. https://doi.org/10.5334/dsj-2021-001

Edmunds, R., L'Hours, H., Rickards, L., Trilsbeek, P., Vardigan, M., & Mokrane, M. (2016). Core trustworthy data repositories requirements. Zenodo. https://doi.org/10.5281/zenodo.168411.

European Union (2024). General Data Protection Regulation: GDPR. Available: https://gdpr-info.eu/

Frank, R. D., Chen, Z., Crawford, E., Suzuka, K., & Yakel, E. (2017). Trust in qualitative data repositories. Proceedings of the Association for Information Science and Technology, 54(1), 102-111. https://doi.org/10.1002/pra2.2017.14505401012

Gualo, F., Rodriguez, M., Verdugo, J., Caballero, I., & Piattini, M. (2021). Data quality certification using ISO/IEC 25012: Industrial experiences. Journal of Systems and Software, 176, 110938. https://doi.org/10.1016/j.jss.2021.110938

Li, J., Krohn, M., Mazieres, D., & Shasha, D. (2004). SUNDR: Secure untrusted data repository. USENIX Symposium on Operating Systems Design and Implementation, 4, 9-25.

Liaw, S. T., Guo, J. G. N., Ansari, S., Jonnagaddala, J., Godinho, M. A., Borelli, A. J., de Lusignan, S., Capurro, D., Liyanage, H., Bhattal, N., Bennett, V., Chan, J., & Kahn, M. G. (2021). Quality assessment of real-world data repositories across the data life cycle: a literature review. Journal of the American Medical Informatics Association, 28(7), 1591-1599. https://doi.org/10.1093/jamia/ocaa340

Lin, D., Crabtree, J., Dillo, I. Downs, R. R., Edmunds, R., Giaretta D., De Giusti M., L'Hours, H., Hugo, W., Jenkyns, R., Khodiyar, V., Martone, M. E., Mokrane, M., Navale, V., Petters, J., Sierman, B., Sokolova, D. V., Stockhause, M., & Westbrook J. (2020). The TRUST Principles for digital repositories. Scientific Data, 7(1), 144. https://doi.org/10.1038/s41597-020-0486-7

Mutula, S. M. (2011). Ethics and trust in digital scholarship. The Electronic Library, 29(2), 261-276. https://doi.org/10.1108/02640471111125212

Oliver, G. & Harvey, D. R. (2016). Digital Curation (2nd ed.). Chicago: American Library Association.

Paprica, P. A., Crichlow, M., Curtis Maillet, D., Kesselring, S., Pow, C., Scarnecchia, T. P., Schull, M. J., Cartagena , R. G., Cumyn, A., Dostmohammad, S., Elliston, K. O., Greiver, M., Hawn Nelson, A., Hill, S. L., Isaranuwatchai, W., Loukipoudis, E., McDonald, J. T., McLaughlin, J. R., Rabinowitz, A., Razak, F., Verhulst, S. G., Verma, A. A., Victor, J. C., Young, A., Yu, J., & McGrail, K. (2023). Essential requirements for the governance and management of data trusts, data repositories, and other data collaborations. International Journal of Population Data Science, 8(4). https://doi.org/10.23889/ijpds.v8i4.2142

Prieto, A. G. (2009). From conceptual to perceptual reality: Trust in digital repositories. Emerald Group Publishing Limited, 58(8), 593-606. https://doi.org/10.1108/00242530910987082

Stvilia, B. & Lee, D. J. (2024). Data quality assurance in research data repositories: A theory-guided exploration and model. Journal of Documentation. https://doi.org/10.1108/JD-09-2023-0177

Trisovic, A., Mika, K., Boyd, C., Feger, S., & Crosas, M. (2021). Repository approaches to improving the quality of shared data and code. Data, 6(2), 15. https://doi.org/10.3390/data6020015

Volentine, R., Owens, A., Tenopir, C., & Frame, M. (2015). Usability testing to improve research data services. Qualitative and Quantitative Methods in Libraries, 4(1), 59-68.

Xafis, V. & Labude, MK. (2019). Openness in big data and data repositories: The application of an ethics framework for big data in health and research. Asian Bioethics Review, 11(3), 255-273. https://doi.org/10.1007/s41649-019-00097-z

Yakel, E., Faniel, I. M., Kriesberg, A., & Yoon, A. (2013). Trust in Digital Repositories. International Journal of Digital Curation, 8(1), 143-156. https://doi.org/10.2218/ijdc.v8i1.251

Yoon, A. (2014). End users' trust in data repositories: Definition and influences on trust development. Archival Science, 14(1), 17-34. https://doi.org/10.1007/s10502-013-9207-8

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

Kim, Juseop, Yang, Seong Jun, & Kim, Suntae (2022). Study on data repository certification scheme analysis and certification strategy: Focused on Coretrustseal. Korean Society for Library and Information Science, 56(2), 209-229. https://doi.org/10.4275/KSLIS.2022.56.2.209