Journal of KIIT. Vol. 22, No. 3, pp. 127-133, Mar. 31, 2024. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2024.22.3.127

바이오 연구데이터 활용서비스 지원방향 도출에 관한 연구

이용호*, 박정우**

A Study on Establishing the Direction of Support for Bio-research Data Utilization Services

Yongho Lee*, Jung Woo Park**

이 논문은 과학기술정보통신부 재원으로 수행된 「데이터 스테이션 구축 및 운영(플랫폼 구축 및 툴 개발) (N24NM016-24)」의 연구비 지원을 받아 수행된 연구임.

요 약

최근 들어 바이오 연구데이터의 축적 및 활용에 대한 국가적 중요성이 높아지고 있다. 이를 지원하기 위하여 공공영역의 체계적인 활용지원 모델 및 서비스 체계가 필요함에도 불구하고 지원 서비스의 개념 및 방향성에 대한 연구는 단편적으로 제시되어 통합 활용 서비스의 개념 설계가 부족한 실정이다. 본 연구는 활용 지원 서비스 개념 도출을 위한 바이오 연구데이터 활용 전주기 정의 및 공공 서비스 분류에 따라 미비한 서비스 개념을 종합적으로 제시한다. 제안한 서비스 개념은 총 8가지로 연구계획단계의 지원, 데이터 품질 평가/확인 단계 지원, 학습 데이터 변환 지원, 데이터 익명화 지원, 데이터 보강/보완 지원, 데이터 공유 지원, 데이터 파기 지원에 대한 세부적 개념 정의 및 절차에 대해 서술하였다. 이를 통해 향후 활용 지원 서비스 시스템 개념 설계에 기여할 수 있을 것으로 생각된다.

Abstract

The importance of gathering and applying bio-research data has grown at the national level in recent years. A systematic utilization support model and public sector service system need to be established to facilitate this. However, research on the concept and direction of support services is presented in a fragmentary manner, resulting in a lack of conceptual design for integrated utilization services. This study comprehensively suggests the service concept that is lacking in accordance with the description of the bio-research data life cycle and the public service category in order to generate the utilization support service idea. Eight service concepts are suggested, and the detailed concept definitions and procedures for supporting the research planning stage, training data conversion, data anonymization, data sharing, data destruction, and data quality evaluation/confirmation stage support are all described. By doing this, it could be possible to contribute to the conceptual design of the utilization support service system in the future.

Keywords
utilization support, public sector service, bio-research data, data life cycle

* 한국과학기술정보연구원 책임연구원
- ORCID: http://orcid.org/0000-0003-0226-4977
** 한국과학기술정보연구원 선임연구원(교신저자)
- ORCID: http://orcid.org/0000-0001-8230-1787

· Received: Dec. 12, 2023, Revised: Jan. 24, 2024, Accepted: Jan. 27, 2024
· Corresponding Author: Jung Woo Park
Korea Institute of Science and Technology Information
Tel.: +82-42-869-0817, Email: j.w.park@kisti.re.kr

Ⅰ. 서 론

최근 4차 산업혁명 시대의 도래로 인해 디지털 기술의 발전과 데이터의 중요성이 대두되면서, 바이오 분야에서도 기존 실험 및 이론 중심의 연구개발에서 데이터 중심의 연구 패러다임으로 전환이 꾸준히 이루어지고 있다.

데이터 활용 및 연구 활성화를 모색하기 위해 정부 주도의 정책이 추진되고 있는데[1]-[3], 제반 데이터 생태계 조성을 위해서 데이터 전 주기(수집·가공·연계·개방·활용) 활성화 및 데이터 플랫폼 구축으로 요약할 수 있다. 구체적으로 데이터 수집과 생성에서 시장 수요에 부합하는 양질의 전문 데이터 공급, 정보 비대칭을 해소하기 위한 지원 플랫폼의 제공, 데이터 활용 사각지대 해소를 위한 주체별 데이터 활용역량 육성 및 활용·분석 인프라 제공이 주요 목표로 추진되고 있다.

바이오 분야에서는 미국, 유럽, 일본 중심의 바이오 데이터 패권 현상을 타개하기 위해 국가 전략 수립[2][3]과 정부 주도의 플랫폼(K-BDS) 구축이 이루어지고 있다.

본 논문에서는 국가 차원의 바이오 데이터 플랫폼 추진 현황을 살펴보고 바이오 데이터 특성과 선도국의 플랫폼 구축 방향을 바탕으로 데이터 분석 및 활용 분야의 플랫폼 구축 및 설계방안을 제시한다. 본 연구를 통해 바이오 연구 데이터의 활용과 공유 활성화에 기여하고, 실제 설계에 반영하여 바이오 데이터 기반 연구 생태계 활성화를 도모할 수 있을 것으로 기대된다.

본 논문의 구성은 다음과 같다. 먼저 국내외 바이오 데이터 활용 플랫폼 관련 정책 및 구축 사례를 조사한다. 다음으로 데이터 활용 활성화를 위한 분석의 틀을 설계하고 이를 기존 사례 및 분석틀과 비교하여 바이오 연구데이터 활용 플랫폼의 구축 방향을 제시한다. 마지막으로 설계 내용에 대한 시사점 및 향후 연구 방향을 제시한다.

Ⅱ. 바이오 연구데이터 활용 플랫폼 분석

2.1 바이오 연구데이터 특성

활용 플랫폼 구축 방향 수립을 위해서는 바이오 연구데이터 활용에 대한 개념 정의가 선행되어야 한다. 바이오 연구데이터는 실험, 관찰, 조사, 분석 등 바이오 R&D 과정을 통해 생산되어 성과 도출에 활용되는 모든 객관적인 사실 데이터로 바이오 연구 수행을 통해 생산, 활용되는 모든 데이터로써 오믹스, 임상·전임상 데이터, 구조데이터, 화합물 분석 데이터 등이 이에 속한다.

또한 데이터 활용이란 데이터 생애주기 관점에서 수집 및 품질검증 단계를 거친 데이터가 새로운 학술적/산업 경제적 목적하에 시각화, 모델링 등 데이터 분석을 수행하여 성과 도출에 기여하는 것을 뜻한다.

데이터 분석 등 활용행태에 영향을 미칠 수 있는 바이오 데이터 특성은 다음과 같다.

첫째, 데이터 기본 크기 및 차원이 높은 특성이 있다. 유전체 데이터 등 주요 바이오 연구 데이터는 대용량 데이터로 생산, 보관된다. 인간 게놈 데이터는 30억 개의 염기쌍, 약 2만 개의 유전자로 구성되어 있다. 단위 샘플당 전장 유전체 약 120GB, 전사체 10GB, 메타지놈 20GB가 생산된다(유전체 원본 시퀀싱 파일(FASTQ) 파일 및 맵핑(BAM) 파일 크기의 총합). 유전체, 단백질, 대사체 등은 데이터 차원이 높아 수백만~수십억 개의 관측치 및 수천~수백만 개의 특성(Feature)을 가질 수 있다. 따라서 대용량 데이터 특성과 차원을 고려하는 저장소 기능이 반영될 필요가 있다.

둘째, 데이터 유형의 다양성이 존재한다. 바이오 연구데이터는 분석 대상에 따라 다양한 형태의 정보가 수집된다. 개인 단위 샘플을 기준으로 볼 때 임상 정보, 유전체·오믹스(단백체, 대사체, 전사체) 데이터, 생체 신호, 의료 영상 등 다양한 데이터가 생산되며 따라서 데이터 표준화 및 통합 분석 이슈가 제기될 수 있다.

셋째, 높은 정보보호 수준이 요구된다. 바이오 분야의 연구데이터 중 인체 유래물 데이터는 개인정보를 포함할 수 있다. 따라서 데이터 보안을 위한 안전한 데이터 저장, 처리 시스템과 보안 프로토콜이 필요하다.

넷째, 데이터 신뢰성을 담보할 수 있어야 한다.

바이오 연구데이터에는 임상 의료분야에 직간접적으로 활용되는 데이터가 있으므로 통상 연구에서 요구되는 데이터 신뢰성 보다 높은 수준으로 데이터 위변조 및 오염이 되지 않았음을 증명할 수 있는 신뢰성 여부가 매우 중요하다. 예를 들어 소유자 정보, 생산자 정보, 샘플 출처를 포함한 샘플 메타 정보, 실험장비 정보, 실험조건 및 데이터 도출 절차 등의 부가적 데이터가 활용할 데이터에 동반되어야 한다.

마지막으로 데이터 이동으로 소유 주체가 변경되는 경우 다양한 법적 제약을 고민하여야 한다. 생명윤리및안전에관한법률, 의료법, 보건의료기본법에 따르면 인체 유래물 기반의 바이오 데이터는 민감정보로 취급하고 있다. 개정된 개인정보보호법에 의거 가명 정보 처리 특례로 데이터를 활용할 수 있는 근거는 마련되었으나 인체유래데이터는 생명윤리법을 우선 적용하여 연구/임상/산업 활용시 절차적 제한이 존재한다. 잠재적 활용 가능성, 데이터 잠재수요 및 활용시의 파급효과를 고려할 때 활용 대상에서 배제하는 것은 타당하지 않으므로 법적 규제와 윤리적 문제를 동시에 고려한 활용 방향을 설계하여야 한다.

2.2 국내 바이오 데이터 플랫폼 현황분석

정부 차원에서 인체유래물 등 바이오 연구소재와 바이오 연구 데이터의 확보·관리·활용을 위해서 2021년부터 11개 범부처 협력체계를 통한 다부처 국가생명연구자원 선진화 사업이 추진 중이다. 해당 사업에서 국가 바이오 데이터의 수집·관리 및 활용을 위한 데이터 공유 플랫폼(K-BDS) 구축 등이 추진되고 있다. 데이터 분야에 ’22년 기준 311억 원이 투입되었으며 데이터 확보에 19.9%, 데이터 관리 50.2%, 데이터 활용 29.9%에 투입한 것으로 조사되었다[4]. 바이오 데이터 활용을 위한 플랫폼 서비스는 CPU/GPU 계산을 위한 고성능 컴퓨팅 자원(컴퓨팅인프라 2,432 core, 55개 GPU(A100), 2PB 스토리지)을 구축하여 제공하는 부문과 분석을 지원하기 위해 외부 바이오 데이터베이스와 오픈소스 분석도구를 분석환경에 탑재하여 연구자에게 제공하는 부문(바이오 DB 20종, 분석도구 19종 제공)이 존재한다[5].

2.3 해외 바이오 데이터 플랫폼 분석

미국은 NIH에서 연구비 지원을 받은 과제에서 산출되는 바이오 연구데이터를 NIH 산하 생물정보센터(NCBI)에 등록을 의무화하고 데이터 타입별 다양한 데이터 등록·저장소를 운영 중이다(ex. SRA, GEO, PubChem등). 최근 바이오 빅데이터와 관련하여 백악관 과학기술정책실(OSTP) 및 상무부(DOC) 주도로 2023년에 5년 내 데이터 표준, 도구 및 능력의 발전과 통합을 통해 효과적이고 안전한 데이터 공유 매커니즘을 포함한 데이터 인프라 구축과 20년 내 생물 제조 제품과 공정의 신속한 개발과 배포를 가능하게 하는 표준 인프라의 구축 계획을 발표하였다[6]. 현재 3,456개 노드, 91,024 CPU core, 35PB 스토리지를 제공하는 인프라를 보유 중이며, AWS와도 제휴하여 관련 도구 및 데이터를 제공하고 있다.

유럽생물정보연구소(EBI)는 NGS 유전체 데이터 ENA, 전사체 데이터 Array Express, 화합물 데이터 chEMBL 등의 데이터 등록 ·저장소를 운영하고 있다. 플랫폼 관련해서는 ELIXIR(European Infrastructure for Life-Science Information)를 통해 6개의 플랫폼을 구성하고 80,000개의 CPU core 및 50PB 스토리지를 연구자에게 제공한다. 6개의 플랫폼은 ① 바이오 데이터의 저장 및 분석을 위한 계산 플랫폼(Compute platform) ②데이터와 문헌 간 연결을 지원하는 데이터 플랫폼(Data platform), ③데이터 분석도구의 저장 및 검색을 지원하는 도구 플랫폼(Tool platform), ④ 유럽 전역의 바이오 데이터 상호 운용성을 보장하기 위한 표준을 제공하는 상호운용 플랫폼(Interoperability platform), ⑤교육 제공 플랫폼(Training platform), ⑥데이터/분석도구/표준 피드백 및 교류를 위한 커뮤니티 서비스(Community service)로 구성된다.

일본은 문부과학성 산하에 바이오데이터 총괄관리 컨트롤타워(NBDC)를 설치하고 범부처 차원의 연구데이터 통합관리를 추진 중이다. DNA 데이터 은행(이하 DDBJ)을 통해 15,424 CPU core 및 47PB 스토리지 플랫폼을 제공한다. 사업내용은 ①공공 바이오 데이터베이스 운영을 통한 데이터 수집, ② 슈퍼컴퓨터 시스템의 관리 및 운영, ③데이터베이스 검색 및 분석 서비스, ④교육 및 홍보 기능으로 구성되어 있다.

중국은 과학아카데미(CAS) 산하에 국가유전체데이터센터(NGDC)를 설립하여 유전체 데이터 및 관련 SW의 수집을 확대하고 있다. 8,800개 CPU core, 48PB 스토리지를 제공 중으로 eGPS 클라우드를 통해 온라인 데이터 분석 서비스를 제공한다.

2.4 바이오 연구데이터 활용 프레임워크 설계

벤치마킹 지원 서비스의 내용을 보면, 결국 바이오 연구데이터의 활용 프레임워크는 바이오 연구데이터의 전주기(Life cycle)단계별 필요한 지원 서비스를 묶어 제공하는 총합적 하부구조로 정의내릴 수 있으므로 본 연구에서는 바이오 데이터 관련 지원서비스를 데이터 활용지원 전주기에 따라 분류하고 벤치마킹 대상인 4개의 해외 공공 플랫폼 서비스가 해당 부문을 지원하는지 확인하여 매칭한다. 이를 위하여 문헌 연구를 바탕으로 바이오 연구데이터의 전주기 단계를 정의한다. 또한 바이오 인포매틱스의 바이오 지원 서비스 분류를 정의하고 벤치마킹 결과와 함께 활용 주기상 각 단계의 어느 부분을 지원하고 있는지 매칭한다. 이를 통해 국가차원의 바이오 활용지원 플랫폼의 설계에 있어서 제공 기능의 지원 방향성을 공백 영역 위주로 파악할 수 있을 것이다.

데이터 기반 바이오 연구의 지원 서비스에 대해 L. Dai et al.[7]은 클라우드 컴퓨팅에서의 바이오 인포매틱스 서비스 현황을 분석하여 이를 4개의 대분류와 8개 세분류로 제시하였다. 4개 대분류는 클라우드 서비스의 기본 유형과 동일하며 세부 분류는 바이오 분야 특성을 조합하여 제시하였다. 4개 대분류는 NIST에서 공식화된 클라우드 서비스의 개념인 Daas(Data as a service), SaaS(Software as a Service), Paas(Platform as a Service), Iaas(Infrastructure as a Service)이다. 세부 개념에서는 바이오 분야의 서비스 특징을 결합하였는데 예를 들어 DaaS는 클라우드를 통해 데이터 자체에 접근하는 서비스인 것은 동일하나 바이오 분야에서 공공 목적으로 생산하여 연구자에게 제공하는 대규모 바이오 연구자원 DB(ex. NCBI)에 대한 접근을 서비스로 제공하는 개념이 포함된다. SaaS는 여러 유형의 바이오 데이터 분석에 사용되는 데이터 분석도구를 클라우드로 제공하는 개념에 분석도구의 버전 관리를 포함시켰다.

Pass와 Iaas는 도메인 특성을 합쳐서 제시했을 뿐 클라우드 서비스 정의와 동일하다. PaaS는 연구자가 바이오 데이터 분석도구를 실행하는데 있어서 필요한 컴퓨팅 환경 및 분석을 위한 프로그래밍 등 개발환경 서비스를 의미한다. IaaS에서는 가상화된 자원(CPU/GPU 및 OS)을 할당하여 제공하며 바이오 데이터를 분석하기 위한 계산자원 및 자원을 조합한 VM(Virtual Machine)을 제공한다.

R. L. Grossman et al.[8]은 데이터과학 측면에서 데이터 활용을 위한 지원 서비스를 4개 유형으로 (data science infrastructure and platform services, data science software as a service, data science support services, data commons) 구분하였다. L. Dai et al.의 연구와 비교할 때 분석도구간 연결을 통해 데이터 처리를 수행하는 파이프라인 및 데이터 사이언스 지원 서비스를 강조하였다. 따라서 본 연구에서는 L. Dai et al.의 연구에 나온 8개 세분류에 R. L. Grossman et al. 분류를 포함한 5개 대분류 - 9개 세부 분류를 서비스 유형으로 제시한다.

바이오 부문의 연구데이터의 전주기 흐름은 연구개발을 수행하기 위해 바이오 데이터가 변환, 활용되는 단계별 흐름으로서 컴퓨팅 환경의 데이터 처리 특성 및 연구개발 단계 특성을 반영하는 절차로 구성되어야 한다. 데이터 전주기와 관련하여, J. M. Wing[9]은 데이터 사이언스에서의 데이터 전 주기를 생성(Generation)에서 해석(Interpretation)까지 진행되는 8단계의 절차를 정의한 바 있다. L. Lin et al.[10], X. Yu et al.[11]는 클라우드 컴퓨팅 환경에 대응한 데이터 전 주기를 제시하였다. L. Lin et al.는 데이터 생성에서 데이터 폐기에 이르는 7단계의 절차를 제시하였다[10]. J. M. Wing의 절차와 비교할 때 L. Lin et al.은 데이터 전송과 접근(Access) 절차를 강조하고 사용자 인증, 데이터 암호화 등을 포함하는 보안 프로토콜 하에서의 데이터 접근 단계를 포함하였다. S. Allard[12]는 과학 연구 흐름에서 연구데이터의 전주기를 분석하였으며 연구 데이터 관련 계획 및 설명과 데이터 신뢰성을 위한 절차로서 계획(Plan), 보장(Assure)단계를 포함한 8단계의 절차를 제시하였다. M. E. Arass et al.[13]은 12개의 데이터 전 주기 분석 연구를 기준으로 상황별 의사결정에 적합한 데이터

전주기 흐름을 선택하는 평가 모델로 11단계의 절차를 제시하였다.

본 연구는 과학적 연구의 흐름에서 활용되는 데이터 전주기 절차(8개 절차)에 빅데이터 전주기(6개), 클라우드 컴퓨팅 환경에서의 데이터 전주기 절차(7개 절차)를 재정리하여 13개의 단계 절차를 정의하였다. 기본 전제는 과학적 연구의 흐름에서 활용되는 데이터 전주기 8단계 절차를 기준으로 동일 개념의 단계를 통합하고 빅데이터의 활용 주기상 재사용(Reuse) 단계, 클라우드 컴퓨팅 주기상 데이터 전송(Transmission) 단계, 익명화(Anonymity) 단계, 데이터 보강(Enrichment), 데이터 결합(Integration) 5단계를 추가한 것이다.

이상의 분석 결과로 도출한 바이오 인포매틱스 지원서비스 9개 분류에 4개 공공 플랫폼의 서비스 지원 여부를 확인하여, 지원하는 플랫폼은 그 약자를 표기하였다(N: NCBI, E: EBI, D: DDBJ, G: NGDC). 다음으로 각 서비스 타입이 지원하는 연구데이터 단계와 매칭하였으며 그 결과는 그림 1과 같다.

따라서 새로운 활용 서비스는 기존 지원 서비스에 더해 기존에는 개별 연구자가 수행해오던 단계를 지원하는 새로운 개념의 활용지원 서비스를 추가하는 형태로 설계된다. 공백영역을 중심으로 하는 지원 서비스의 조작적 정의 및 개념적 프로세스를 설계하면 다음과 같다.

(1) Plan 지원: 데이터-AI 기반 가설 도출 지원

해당 서비스는 데이터를 기준으로 데이터 시각화하고 통계분석을 수행하여 패턴, 관계, 이상치를 발견하는 탐색적 데이터 분석(EDA)을 실시한다. 이를 통해 데이터를 기반으로 한 새로운 가설을 도출하도록 지원한다. 예를 들어 특정 약물이 특정 단백질 반응 체계에 미치는 영향에 대한 자동 가설을 생성한다. 자동으로 도출된 가설은 연구자 커뮤니티에 제시되어 평가, 수정되고 이를 통해 새로운 연구가설을 기반으로 한 연구가 착수되도록 지원한다.

(2) Assure 지원: 데이터 품질 평가 및 확인 지원

새로운 데이터 등록시 등록 전 단계에서 데이터 서비스 주체가 별도로 조직한 전문가 커뮤니티가 데이터 품질에 대해 평가하고 그 신뢰성을 확인한다. 공개 데이터에 대해서도 동일한 절차를 수행하여 연구자가 공개 데이터를 활용할 때 개별적으로 데이터 신뢰성을 확인하는 시간과 비용을 절감하도록 한다.

(3) Describe 지원: AI 학습데이터 전환 지원

AI학습 데이터 전환시 기계학습용 데이터로의 전환에 필요한 레이블링 생성 등 메타데이터 생성 지원 및 표준화 지원 서비스를 수행한다.

Service type	description	Service availability (N/E/D/G)
DaaS	Data Storage Service	N/E/D/G
DaaS	Access to Public Datasets	N/E/D/G
SaaS	Providing Tools	N/E/D/G
SaaS	Providing Pipelines	E
PaaS	Analysis platform Service	E/D/G
PaaS	Programming environments service	E/D/G
IaaS	Virtual Machine	E/D/G
IaaS	Virtualized resources	E/D/G
DSSS	Data Science Support Service	E/D

Data lifecycle	description
Plan	Building research plans
Collect	collect public/private dataset
Assure	assure data quality
Describe	discover data and writing metadata
Transmit	data transfer
Store	Storing data
Integrate	Integrating data
Anonymize	Anonymizing data
Enrich	data query/extract/transform
Analyze	Analyzing data
Share(reuse)	Sharing with others
Destruct	data disposal
Archive	permanent archiving data

그림 1. 활용지원서비스분류-벤치마킹결과-바이오데이터활용주기 매칭결과

Fig. 1. Utilization support service classification - benchmarking - matching results for the bio-data utilization lifecycle

(4) Integrate 지원: 데이터 통합 분석 지원

서로 다른 소스에서 수집된 동종, 이종 데이터의 통합 데이터셋 구축을 지원하는 것으로 데이터 표준화를 통한 동일한 형식 변환, 결측치 처리 및 이상치 제거에 대한 반구조화된 자동지원, 고유한 식별자 확인 및 이를 이용한 통합데이터 구축지원 서비스를 수행한다.

(5) Anonymize 지원: 데이터의 익명화·가명화 지원

바이오 데이터의 익명화 및 가명화 지원은 개인정보로 식별되는 데이터의 모든 식별자를 확인하고 이를 무작위 또는 암호화된 식별자로 대체하는 작업을 지원하는 서비스로 정의된다. 개인정보의 일부 분석이 필요한 경우 개별 수치정보를 범주화하는 정보 변환 지원을 포함한다(ex. 나이, 주소 정보의 범주화 등). 필요시 무작위 노이즈를 추가하는 작업을 포함한다.

(6) Enrich 지원: 데이터 보강/보완 지원

원본 데이터셋에 외부 공개 데이터의 연결을 지원하는 서비스로 공공, 상업 데이터베이스의 연결 및 결합을 통한 새로운 파생변수의 생성을 지원한다.

(7) share 지원: 데이터 공유 지원

연구자와 연구자간 협업시 데이터셋의 공유를 위한 필요 서비스 모듈(데이터 스토리지, 특정 대상으로의 제한된 데이터 공유)을 지원한다. 해당 지원을 위해서는 연구자간 협업 수요와 공급 중개서비스가 필요하며 본 지원 모듈은 해당 중개서비스를 포함한다.

(8) Destruct 지원: 데이터 파기 지원

불필요한 데이터 파기 및 파기 기록을 남겨 향후 관리 책임에 대해 근거를 남기고 이를 확인할 수 있도록 지원한다.

III. 결론 및 향후 과제

바이오 데이터 플랫폼은 미국, 유럽, 일본과의 경쟁에서 국가 차원의 전략 수립과 플랫폼 구축이 이루어지고 있다. 본 연구에서는 바이오 연구데이터의 활용을 위해 바이오 연구데이터의 특성을 분석하고 해외 플랫폼의 특징에 대해 벤치마킹 분석을 실시하였다. 기존 문헌 연구를 기반으로 바이오 연구데이터 전주기 및 공공 바이오 인포매틱스 서비스 분류를 제시하고 이를 매칭하여 현재 미비한 단계의 새로운 연구데이터 활용 지원서비스 개념에 대해 제안하였다.

향후 과제로 본 서비스 개념의 실제 설계에 필요한 기능 요소 파악 및 흐름 설계가 필요하며 또한 각 서비스 필요성에 대해 연구자 수요조사가 필요하다. 바이오 분야에서의 데이터 활용과 플랫폼 구축에 대한 본 연구는 국가 차원의 데이터 활용 플랫폼 전략 수립과 기본 설계 방향에 심층적인 이해를 제공하며, 향후 구체적인 시스템 설계 및 서비스 개념 설정에 기여할 수 있을 것으로 기대된다.

References

[1] Ministry of Science and Technology Information and Communication, Multi-Agencies, "Big Data Strategy for Biological Research Resources", Jul. 2020.

[2] Ministry of Science and Technology Information and Communication, Multi-Agencies, "The 3rd National Basic Plan for the Management and Utilization of Life Sciences Research Resources", May. 2020.

[3] Ministry of Science and Technology Information and Communication, "Digital Bio-Innovation Strategy", Dec. 2022.

[4] Ministry of Science and Technology Information and Communication, "2023 National Life Sciences Research Resource Management and Utilization Implementation Plan", Dec. 2022.

[5] KISTI, "2023 annual report: K-BDS Establishment and Operation (Platform Development and Tool Creation)", Oct. 2023.

[6] BPRC, "Executive Orders and Challenges of the Biden Administration in the US", BioINpro, Vol. 112, 2023.

[7] L. Dai, X. Gao, Y. Guo, J. Xiao, and Z. Zhang,

"Bioinformatics clouds for big data manipulation", Biology Direct, Vol. 7, No. 43, Nov. 2012. https://doi.org/10.1186/1745-6150-7-43

[8] R. L. Grossman, A. Heath, M. Murphy, M. Patterson, and W. Wells, "A Case for Data Commons: Toward Data Science as a Service", Computing in Science & Engineering, Vol. 18, No. 5, pp. 10-20, Sep. 2016. https://doi.org/10.1109/MCSE.2016.92.

[9] J. M. Wing, "The Data Life Cycle", Harvard Data Science Review, Vol. 1, No. 1, Jul. 2019. https://doi.org/10.1162/99608f92.e26845b4.

[10] L. Lin, T. Liu, J. Hu, and J. Zhang, "A privacy-aware cloud service selection method toward data life-cycle", 2014 20th IEEE International Conference on Parallel and Distributed Systems (ICPADS), Hsinchu, Taiwan, pp. 752-759, Dec. 2014. https://doi.org/10.1109/PADSW.2014.7097878.

[11] X. Yu and Q. Wen, "A view about cloud data security from data lifecycle", 2010 International Conference on Computational Intelligence and Software Engineering, Wuhan, China, pp. 1-4, Dec. 2010. https://doi.org/10.1109/CISE.2010.5676895.

[12] S. Allard, "Dataone : Facilitating escience through collaboration", Journal of eScience Librarianship, Vol. 1, No. 1, pp. 3, Feb. 2012. https://doi.org/10.7191/jeslib.2012.1004.

[13] M. E. Arass, I. Tikito, and N. Souissi, "Data lifecycles analysis: Towards intelligent cycle", 2017 Intelligent Systems and Computer Vision (ISCV), Fez, Morocco, pp. 1-8, Apr. 2017. https://doi.org/10.1109/ISACV.2017.8054938.

저자소개

이 용 호 (Yong-Ho Lee)

1999년 : KAIST 산업경영학과 (공학사)
2001년 : 서울대학교 협동과정 기술경영전공(공학석사)
2013년 : 서울대학교 산업공학과 (공학박사)
2003년 ~ 현재 : 한국과학기술정보연구원 책임연구원
관심분야 : 기술혁신, 기술정보분석, 과학기술 및 바이오정책연구

박 정 우 (Jung Woo Park)

2008년 : 이화여자대학교 화학과 (학사)
2014년 : POSTECH 화학과 이학박사(석박사통합과정)
2015년 ~ 현재 : 한국과학기술정보연구원 선임연구원
관심분야 : 유전체분석, 바이오빅데이터 분석, 과학기술 및 바이오정책연구