https://doi.org/10.5392/JKCA.2017.17.12.117

국내 과학기술분야 연구기관의 과학데이터 관리 현황

Research Data Management of Science and Technology Research Institutes in Korea

최명석, 이승복, 이상환
한국과학기술정보연구원 과학데이터연구센터

Myung-Seok Choi(mschoi@kisti.re.kr), Seung-Bock Lee(sblee@kisti.re.kr), Sanghwan Lee(sanglee@kisti.re.kr)

요약

최근 연구 환경과 연구 패러다임이 데이터 중심(Data-Driven)으로 변화되고 있다. 특히, 공공 연구성과의 개방과 공유에 기반한 오픈 사이언스(Open Science)가 과학 연구의 글로벌 어젠다로 새롭게 부각되고 있다. 해외에서는 공적 지원에 의해 수행된 연구에서 생산되는 과학데이터의 공유·활용을 위한 정책이 적극적으로 시행되고 있어, 국내에서도 '오픈 연구데이터'를 위한 정책과 인프라 구축이 시급한 상황이다. 본 연구에서는 국내 과학기술 분야 연구기관의 과학데이터 생산, 관리, 활용 현황을 조사했다. 국가과학기술연구회 소속 22개 정부출연 연구기관과 국내 20개 대학의 연구자를 대상으로 과학데이터 생산, 관리, 활용 현황, 과학데이터 공유·활용 참여 의지, 과학데이터 공유·활용 요구사항 등 5개 관점의 심층인터뷰를 실시했다. 이를 기반으로 과학데이터 공유·활용을 위한 시사점과 개선방향을 도출했다.

■ 중심어 : | 과학데이터관리 | 연구데이터관리 | 오픈사이언스 | 오픈연구데이터 | 데이터관리계획 |

Abstract

As the recent research environment and research paradigm have become data-driven, Open Science, based on openness and sharing of public research results, has emerged as a global agenda for scientific research. National policies for sharing and re-use of research data from publicly-funded research are in effect globally. Therefore, in Korea, it is urgent to build policies and infrastructure for sharing and re-use of research data. In this paper, we investigate the current status of research data management of science and technology research institutes in Korea. We conducted in-depth interviews with researchers from 22 research institutes belonging to the National Research Council of Science & Technology, and 20 universities in Korea, asking about terms of creation·management·utilization of research data, willingness to share data, and needs for sharing and re-use of research data. From these interviews, we drew implications for open research data and future directions.

■ keyword : | Scientific Data Management | Research Data Management | Open Science | Open Research Data | Data Management Plan |

* 본 연구는 한국과학기술정보연구원(KISTI) 주요사업 과제로 수행된 연구입니다.
접수일자 : 2017년 09월 08일
수정일자 : 2017년 10월 19일
심사완료일 : 2017년 10월 19일
교신저자 : 이상환, e-mail : sanglee@kisti.re.kr

I. 서 론

첨단연구장비, 센서, 데이터 처리 기술 등 디지털 기술의 발달로 대용량 과학데이터가 폭발적으로 생산되면서 데이터 중심의 4세대 연구 패러다임이 등장했다[1]. 이와 더불어 공유·융합을 강조하는 R&D 흐름이 공적 지원에 의해 생산된 연구결과물의 쉬운 접근과 활용을 위한 오픈 사이언스 패러다임으로 확대되고 있다. OECD는 오픈 사이언스를 정책의제로 채택하고 있으며, 세계 각국에서도 오픈 사이언스 확산을 위한 정책을 추진하고 기반 인프라 구축에 주력하고 있다[2-5].

오픈 사이언스 운동은 최근 연구 출판물의 오픈 액세스에서 오픈 연구데이터로 확장되고 있다. 오픈 연구데이터는 연구 과정에서 생산된 과학데이터[1]에 대한 자유로운 접근 및 재사용을 허용하는 것으로[2], 데이터 중심 연구와 전 지구적 문제해결을 위한 협동연구 활성화의 촉매 역할을 하고 있다. 주요 선진국에서는 공공자금이 투입된 연구과제로부터 생산된 과학데이터의 체계적인 관리와 쉬운 접근, 재사용을 통한 가치 창출을 위해 데이터 관리 계획(Data Management Plan, DMP)를 비롯한 오픈 연구데이터 정책을 시행하고 있다[6-10].

개방(Openness)은 과학 연구의 중요한 특성인 자체 교정(Self-correction)이 작동하기 위한 기본 전제이며, 과학적 발견을 촉진함과 동시에 연구의 투명성을 확보함으로써 나쁜 과학(Bad Science)을 찾아내고 근절하는데 도움을 주는, 과학 연구의 본질적 기제에 속한다고 볼 수 있다[3]. 사회경제적으로도 유럽 생물정보학 연구소(EMBL-EBI)의 과학데이터 공유·재활용을 통한 경제적 이득은 연간 1조 파운드, EMBL-EBI 데이터 및 서비스를 활용한 R&D의 ROI는 연간 9.2억 파운드로 추정하고 있으며[11], 데이터센터의 과학데이터 구축·공개로 인해 투자 대비 2~10배 정도의 수익을 창출할 것으로 예상되고 있다[12]. 또한, 공유 인프라 기반의 빠른 지식 확산을 통해 연구자뿐 아니라 Citizen Science 등 기업과 사회 구성원에게도 편익을 제공할 수 있게 된다.

최근 연구 재현성(Reproducibility)에 대한 위기의식이 높아지고 있다[13]. 유망한 종양학 논문 53개 중 47개 연구가, 생명의학 논문 중 50% 이상이 재현 불가능하며[14][15], 2008년 심리학 분야의 저명 학술지 3개에 출판된 100개 실험에 대해서는 약 1/3~1/2만이 재현 가능했다[16]. 또한 세계적 유명 학술지에 논문과 함께 제출된 데이터에서 많은 오류가 발견되고 있으며, 해마다 증가하고 있다[17]. 이에 따라 기존 연구의 재현성 검증을 수행하는 전문 저널이나 기관이 생겨나고 있으며(Preclinical Reproducibility and Robustness[2], Center for Open Science[3], Metrics[4] 등), 검증이 제대로 이루어지지 않을 경우 논문 철회로도 이어지고 있다. 연구 재현성 검증을 위해서도 연구에서 생산되고 사용된 데이터의 공개가 필수적이라 할 수 있다.

과학데이터의 국가적 관리와 활용은 급변하는 미래를 대비하기 위한 글로벌 어젠다이자, 연구 투명성 제고 및 효율화를 위한 필수 요소이다. 국내의 경우 국가 R&D정보는 NTIS[5]를 통해 통합관리 중이며, 공공기관이 보유하고 있는 공공데이터와 일부 과학데이터에 대한 공유·활용 제도가 마련되어 있으나, 다양한 유형의 과학데이터를 국가연구개발 사업의 결과물로 인정하고 국가 차원에서 공유·활용하기 위한 법적 기반과 관련 인프라가 아직 미흡한 실정이다. 본 논문에서는 국내 과학기술 분야 연구기관의 데이터 관리·활용 현황을 살펴보고자 한다.

본 논문의 구성은 다음과 같다. 먼저 데이터 관리·활용 현황 조사와 관련된 기존 연구에 대해서 살펴본다. 그리고 2015년과 2016년에 걸쳐 수행된 과학기술 분야 정부출연 연구기관과 주요 대학의 현황조사 결과와 시사점에 대해 알아본다. 다음으로 국가 차원의 과학데이터 관리·활용체계 마련을 위한 개선 방향을 제시한다. 마지막으로 연구 결과를 종합하며 결론을 맺는다.

1 본 논문에서 과학데이터는 연구 활동의 과정 또는 결과로 산출되는 데이터로, 연구데이터와 동일한 의미로 사용하였음.

2 https://f1000research.com/gateways/PRR

3 https://cos.io

4 https://metrics.standford.edu

5 http://www.ntis.go.kr

II. 관련 연구

Tenopir의 연구에서는 2014년 미국과 캐나다의 ACRL (Association of College & Research Libraries) 소속 도서관 책임자 128명을 대상으로 도서관에서의 연구데이터서비스(Research Data Service, RDS)에 대한 설문조사를 실시했다[18]. 아직 상당수의 도서관들이 RDS를 제공하고 있지 않은 상황이지만, 도서관의 데이터 큐레이션 참여에는 공감대를 이루고 있었으며, 데이터 검색 및 인용, 데이터 활용 가이드라인 등에 대한 요구가 가장 많았다.

Barsky의 연구에서는 캐나다의 UBC(University of British Columbia) 소속 연구자 100명을 대상으로 데이터 관리 현황에 대한 설문조사를 실시했다[19]. 연구자들은 로컬 하드디스크를 주로 활용하고 있었으며, 외부의 데이터 리포지터리를 활용하는 경우는 19.6%에 불과했다. 데이터 공유를 전혀 하지 않는 비율은 12.6%이었으며, 대부분은 개인의 인적 네트워크를 통해 공유하고 있었다. 하지만 79.4%의 연구자들이 데이터 공유가 재현가능하고 협력적인 과학을 강화할 것이라는데 동의했으며, DMP 작성과 데이터 리포지터리, 데이터 관리 지원을 가장 필요한 서비스로 꼽았다.

Shearer의 연구에서는 2016년 12월 COAR(Confederation of Open Access Repositories)에 가입한 리포지터리 관리자 43명을 대상으로 연구데이터 관리 니즈에 대한 설문조사를 실시했다[20]. 53%가 이미 데이터를 수집하고 있었으며, 그렇지 않은 응답자의 72%도 가까운 미래에 데이터 수집 계획이 있다고 응답했다. 응답자의 50%가 데이터와 출판물에 대해 동일한 리포지터리를 활용하고 있었으며 DSpace와 Dataverse가 가장 많이 사용되는 플랫폼이었다. 연구데이터 수집에 있어 가장 어려운 점은 연구자 참여, 기관의 연구데이터 정책 부족, 데이터 저장과 보존을 위한 인프라스트럭처 등이었다.

김문정의 연구에서는 과학기술 분야 연구자 198명을 대상으로 연구데이터 공유에 대한 주요 영향 요인 분석을 위한 설문조사를 실시했다[21]. 보상체계는 데이터 공유에 유의미한 영향을 미치며, 인지성, 의사소통의 개방성, 신뢰성, 협력성 중 인지성만이 이러한 보상체계를 통해 데이터 공유에 긍정적인 영향을 끼치는 것으로 나타났다.

김지현의 연구에서는 대학 연구자 13명의 인터뷰를 통해 연구데이터의 관리와 공유에 대한 인식을 조사했다[22]. 연구자들은 다양한 유형과 포맷의 데이터를 생성·수집하고 있었으며 데이터 문서화를 수행하는 연구자들은 소수에 불과했으나 그 중요성은 인식하고 있었다. 데이터의 공유와 재사용은 개인적인 연구 네트워크의 범위 내에서 이루어지고 있었다. 다수의 연구자들이 아이디어 도용, 표절, 논문 출판의 주도권 문제 등 데이터 공유에 대해 우려하고 있었으며 이를 위한 유인책이 마련될 필요가 있다고 지적했다.

선행 연구에서는 주로 특정 단체의 소속원들에 대한 설문 조사를 통해 전반적인 데이터 공유 인식, 현황 등에 대해 살펴보았다. 본 연구에서는 국가 차원의 데이터 공유·활용 정책 수립을 위해 정부출연 연구기관과 대학을 포괄하는 국내 주요 과학기술 분야 연구기관을 대상으로 구체적인 데이터 관리·활용 현황 및 요구사항을 조사하고 개선방향을 도출했다.

III. 정부출연 연구기관 및 대학 현황조사 결과

1. 현황조사 개요

과학데이터 관리·활용 현황조사는 국가R&D에서 가장 많은 비중을 차지하는 정부출연 연구기관(이하 출연연)과 대학을 대상으로 2015년과 2016년에 걸쳐 수행되었다.

- 목적 : 과학데이터 공유·활용 추진을 위한 데이터 생산·관리·활용 현황 파악 및 개선점 도출
- 대상 및 조사기간 : 과학기술분야 22개 정부출연 연구기관(2015년) 및 20개 대학(2016년)[6]
- 조사방법 : 설문조사 및 심층인터뷰
- 조사내용 : 데이터 생산·관리·활용 현황, 데이터 공유·활용 참여 의지 및 요구사항

6 연구기관별로 데이터를 생산·활용하는 핵심과제를 선정하여 출연연 22개 과제, 대학 11개 과제의 연구책임자 및 데이터 관리자를 대상으로 심층인터뷰를 진행했으며, 대학의 경우 연구자 외에 데이터 관리에 관련된 도서관과 산학협력단에 대한 조사도 실시했다.

2. 현황조사 결과

2.1 생산 현황

먼저 국가R&D과제에서 재사용 가능한 다양한 유형의 데이터가 생산되고 있음을 확인했다. 과학데이터는 장비·실험으로부터 관측·측정을 통해 생산되는 경우가 가장 많았으며, 이미지, 동영상, 텍스트 등 비정형 데이터도 상당한 비중을 차지하고 있었다. 출연연의 R&D 과제에서 생산되는 데이터를 [그림 1]과 같이 분류할 수 있다. 가로축으로는 데이터의 활용 형태에 따라 원천 데이터(Raw 또는 가공)의 형태로 활용되는 유형 A, 가공·분석을 통해 참조 데이터로 활용되는 유형 B, 연구의 해석 및 분석 결과를 담고 있는 유형 C로 나누고, 세로축으로는 데이터 생산 형태에 따라 데이터 확보가 목적인 과제에서 생산되는 유형 가와 전체 연구 공정에서 부가적으로 데이터가 생산되는 유형 나로 분류했다.

그림 1. 출연연 과학데이터 유형

과학데이터 생산이 목적인 과제에서 국가 차원에서 공유·활용 가능한 데이터(영역 I, II)가 다수 생산되고 있었다. 활용을 목적으로 데이터를 생산하기 때문에 표준화 및 품질 검증이 비교적 잘 되어 있고, 관리수준도 상대적으로 높은 편이었다. 하지만, 데이터 생산에 대한 노력이 연구 성과로 인정받지 못하고 과제에 대한 연속성이 보장받지 못하는 등 데이터를 안정적으로 생산하고, 장기적인 관리·활용 계획을 수립·시행할 수 있는 지원이 부족한 상태로 확인되었다.

영역 III은 연구의 각 단계에서 부가적으로 생산되는 데이터로, 연구자들이 데이터를 자신의 연구 자산으로 여기는 경우가 많으며, 이는 공유·활용에 가장 큰 저항감으로 이어지고 있었다. 따라서 이 영역은 데이터 관리 의무를 강조하면서, 장기적인 관점에서 데이터 공유 문화 정착을 유도하는 접근이 필요하다.

연구의 해석 및 분석 결과를 담고 있는 영역 IV는 주로 논문과 보고서에서 표나 그림의 형태로 압축적으로 제시되는 데이터로, 최근 많은 학술 저널에서 논문의 증빙자료로서 Supplementary Data의 형태로 제출을 요구하는 경우가 많다.

대학의 경우 논문, 특허를 위한 소규모 연구 과제를 수행하는 경우가 많으며, 따라서 연구과정에서 부가적으로 데이터를 생산하는 경우가 대부분이었다.

그림 2. 대학 과학데이터 유형

과학데이터의 공유는 논문, 보고서 등의 다른 연구 성과물과는 달리 재사용을 염두에 두고 활용 관점에서 접근해야 하며, [그림 1]의 I, II, III 영역의 데이터에 보다 중점을 두고 추진되어야 한다. 또한, 효율적인 공유·활용을 위해서는 데이터 특성에 따라 단계적인 접근이 필요하다. 데이터 생산 시 실험조건, 기법, 절차 등 생산 환경에 대한 정보를 담고 있는 프로토콜, 프로비넌스 데이터와 데이터를 처리·분석하기 위한 소프트웨어도 데이터 해석 및 활용에 필수적이며, 데이터와 함께 접근·활용이 가능해야 한다.

2.2 관리 현황

생산된 과학데이터는 개인·부서 차원에서 PC나 외장하드를 활용하여 단순하게 저장·관리하고 있는 경우가 대부분이었다.

그림 3. 출연연 과학데이터 관리 현황

출연연의 경우 체계적인 데이터 관리 규정과 전용 데이터 관리 시스템을 갖추고 있는 경우는 거의 없었으며, 일부 과제에서만 과제 단위의 데이터 관리 시스템을 운영하고 있었다. 또한, 과학데이터 공유·활용을 위해서는 개별 기관의 관리수준 향상이 필수적이나, 국가 차원의 표준 가이드라인, 전문 인력 등의 지원이 전무했다.

연구자들은 데이터 등록, 관리, Q&A 대응 등 데이터 관리를 추가적인 업무로 인식하며 이에 대한 업무 부담이 매우 높았다. 특히, 데이터 공개 이후의 품질 및 신뢰성 책임 이슈에 민감했다.

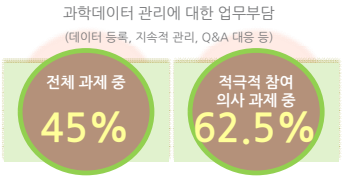


그림 4. 출연연 과학데이터 관리 부담

대학의 경우 주로 석박사 과정 대학원생으로 구성된 연구 인력의 잦은 교체로 인해 데이터 관리 전문성을 갖추기가 더욱 어려운 형편이었다. 또한 도서관과 IT 지원 부서를 통해 데이터 관리를 지원하고 있는 해외의 많은 대학의 경우와 달리 국내 대학의 도서관에서는 데이터 관리 지원에 대해 아직 엄두를 내고 있지 못하는 상황이다.

그림 5. 대학 과학데이터 관리 현황 (1)

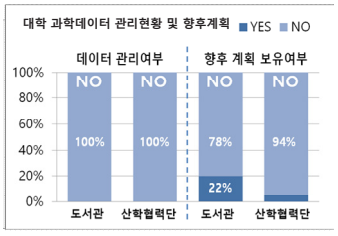


그림 6. 대학 과학데이터 관리 현황 (2)

과학데이터의 활용에 있어 연구자들이 가장 중요하게 인식하는 요소는 데이터의 품질이었다. 연구자들은 자체적인 품질 기준을 가지고 데이터 품질 관리를 수행하고 있었고, 외부의 데이터를 활용할 때도 데이터 품질 보증을 요구하고 자체적으로 품질을 검증한 후 연구에 활용하고 있었다. 하지만 이러한 데이터 품질 관리 수준은 연구 분야별 또는 연구자별로 그 방법과 절차·기준이 상이하여 실제 품질 수준에 대해서는 객관적으로 판단하기 어렵다. 또한 앞서 살펴본 바와 같이 최근 연구 재현성의 문제가 대두되면서 학술 저널에 제출되는 데이터에 문제가 빈번하게 발견되고 있다. 이는 연구자 혹은 연구실 수준의 개별적인 데이터 품질 관리로는 한계가 있음을 보여주며, 연구 신뢰성 확보를 위한 보다 체계적인 품질 검증 체계가 필요하다.

그림 7. 과학데이터 품질 관리 현황

2.3 활용 현황

연구자들은 주로 개인의 인적 네트워크를 통한 요청에 의해 생산된 과학데이터를 공유하고 있었다. 특히 대부분 단독 과제를 중심으로 연구를 수행하는 대학 연구자들은 데이터 공유 필요성을 크게 느끼지 않고 있었으며, 전반적인 공유 인식 또한 높지 않은 것으로 나타났다.

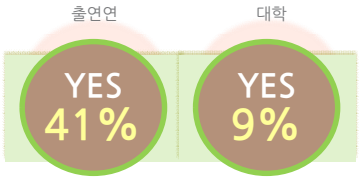


그림 8. 과학데이터 활용 현황

하지만, 연구 커뮤니티 혹은 국가 차원의 과학데이터 공유·활용에 대한 필요성은 인지하고 있었다. 출연연 답변자의 36%가 적극적 참여, 55%가 제도 및 플랫폼

등의 환경 구비 조건으로 참여의사를 밝혔으나, 정부 주도의 정책 추진에 대해서는 부정적인 인식을 가지고 있었다.

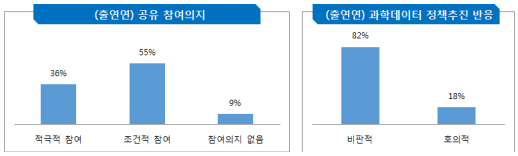


그림 9. 출연연 과학데이터 공유 인식

대학의 경우 연구자들의 55%는 과학데이터 공유 필요성을 인지하고 있었으며 참여의지(45%)도 있으나, 의무화 방식은 반대하는 경우가 많았다(73%).

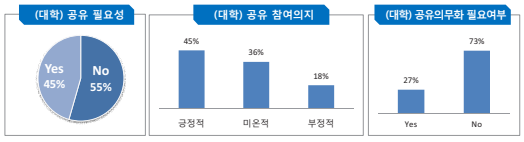


그림 10. 대학 과학데이터 공유 인식

과학데이터 공유·활용을 위한 국내의 데이터 인프라가 미흡한 실정이다. 대학 연구자 중 88%가 해외 리포지토리를 이용하고 있으며 국내 인프라의 부재7, 손쉬운 접근 및 활용, 고품질의 데이터 보유 등의 이유로 해외 리포지토리 이용을 선호하고 있었다.

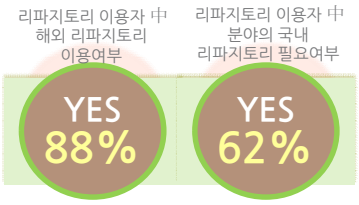


그림 11. 해외 리포지터리 이용 현황

대부분의 연구자들은 생산된 과학데이터를 개인 소유의 연구 자산으로 인식하고 있었으며, 이는 공유에 있어 가장 큰 저항감으로 작용하고 있었다. 따라서 공적 지원에 의한 연구에서 생산되는 데이터는 연구기관의 (공동) 소유로 간주하는 경우가 많은 해외와 같이 소유권에 대한 명확한 법적 기준을 마련한 필요가 있다.

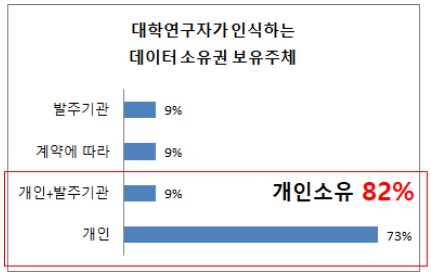


그림 12. 과학데이터 소유권 인식

연구의 핵심 자산인 과학데이터는 아직 연구 성과로 인정받고 있지 못하며, 과학데이터를 주로 생산하는 과제에서도 논문, 특허 등의 성과지표로 평가를 받고 있는 상황이다. 연구자들은 과학데이터의 연구 성과 인정(대학 72.7%), 생산자 권리보호 방안 수립(출연연 100%)을 법제도의 주요 요구사항으로 제시하고 있었다.

IV. 과학데이터 관리 · 활용체계 개선 방향

과학기술 분야 연구기관의 현황 조사를 기반으로 과학데이터 공유·활용을 위한 개선 방향을 도출했다.

1. 데이터 관리 계획 수립 의무화 정책 추진

먼저 공적 지원에 의해 생산된 과학데이터가 효율적으로 관리·공유될 수 있도록 국가R&D과제의 계획서에 DMP 기재를 의무화하는 제도를 수립하고 이를 지원하는 가이드라인 및 인프라 마련이 추진되어야 한다. 미국, 영국, 호주 등 주요 선진국에서는 데이터 관리 계획 의무화 정책을 적극적으로 시행하고 있으며, 연구자들의 체계적인 데이터 관리와 공유 인식 제고를 위한 중요한 요소로 인식하고 있다.

국가연구개발사업의 관리 등에 관한 규정 개정을 통해 국가R&D과제 추진 시 데이터 관리 계획 제출과 평가에 대한 법적 근거를 마련할 필요가 있으며, 연구펀

7 데이터 리포지터리 등록 사이트인 re3data.org에 따르면 2017년 7월 현재 전 세계에서 1,800여개의 리포지터리가 등록되어 있으나, 연구자들이 활용 가능한 국내 리포지터리는 거의 없는 실정이다.

딩기관을 중심으로 정책을 시행하되, 전문지원기관 또는 분야별 데이터센터 등을 통해 연구자의 데이터 관리 계획 작성 및 데이터 관리·공유를 지원할 수 있는 체계도 필요하다.

이를 위해서는 과학기술기본법과 연구성과평가법 등의 개정을 통해 과학데이터를 국가R&D 성과물 및 성과관리 대상으로 지정하고, 과학데이터의 보존·관리·활용에 대한 절차·방법 규정 등 제도적 근거 마련도 동시에 이루어져야 한다.

장기적으로 다양한 분야의 과학데이터를 효과적, 집중적, 체계적으로 관리·활용하기 위해서는 과학데이터의 관리·활용에 대한 독립적인 법률 제정이 효과적인 제도적 방안이다.

법제도, 관리체계, 인프라 관점의 주요 추진 과제는 다음과 같다.

표 1. 주요 추진 과제 (1)

구분	주요 추진 내용
법제도	· 국가 과학데이터 관리 · 공유를 위한 법적 근거 마련 · DMP 제출 의무화 근거 마련
관리체계	· DMP 실행 체계 수립 · 분야별 DMP 작성 가이드라인 마련 및 지원
인프라	· DMP 작성 지원 도구 제공 · 국가 R&D 데이터 현황 및 데이터맵 제공

2. 과학데이터 성과 인정 및 인센티브 방안 마련

연구자들이 데이터 생산과 공유에 대한 필요성을 인식하고 자발적인 의지를 가질 수 있는 제도와 정책이 필요하다. 과학데이터가 논문이나 특허와 같이 공식적인 연구 성과로 인정받으며, 우수한 과학데이터의 생산 및 활용 지원도 연구 성과 평가에 반영될 수 있는 방안을 수립해야 한다. 데이터 인용 등을 통해 데이터를 생산한 연구자에게 학문적 크레딧이 돌아갈 수 있고, 데이터로 인해 금전적 이익 발생시 배분할 수 있도록 과학데이터 공개 및 활용에 따른 연구자의 권익보호 장치를 마련해야 한다. 또한 과학데이터 활용으로 인해 문제가 발생한 경우, 데이터를 생산한 연구자의 면책 조항, 이의신청 근거 및 절차를 명확화할 필요가 있다.

표 2. 주요 추진 과제 (2)

구분	주요 추진 내용
법제도	· 과학데이터를 연구성과로 인정하는 법제도적 근거 마련 · 과학데이터 소유권 및 권리행사 기준 마련 · 연구성과평가 제도 개선
관리체계	· 과학데이터 권리보호 및 인센티브 방안 마련 · 국가 핵심 과학데이터 생산 과제 육성 및 지원 · 과학데이터 출처 표기 필수화 · 과학데이터 활용 가이드라인 제공
인프라	· 데이터 출판 체계 마련 및 활용 모니터링 · 과학데이터 마켓플레이스 환경 지원

3. 개인정보보호 등 데이터 활용 규제 완화 방안 마련

국가R&D과제에서 생산된 과학데이터 중 활용을 제한하는 규제나 법률 때문에 공유가 어려운 데이터가 존재한다. 특히, 인간의 신체·유전 관련 데이터는 많은 생산 비용이 투자되고 재활용 가치가 높으나 개인정보보호법 등에 따라 공유가 매우 어려운 상황이며, 부처 혹은 기관간의 장벽에 의해 데이터 공유가 어려운 경우도 많다. 이러한 각종 데이터 활용 규제를 완화하는 방안을 마련해야 한다.

표 3. 주요 추진 과제 (3)

구분	주요 추진 내용
법제도	· 개인정보 관련 규제 완화 · 데이터 생산부처의 데이터 활용 규제 완화
관리체계	· 과학데이터 개인정보보호 가이드라인 제공 · 부처별 데이터 보유 현황 제공 및 공개 기준 정립
인프라	· 공유 데이터 내 개인식별정보 검증 지원 · 개인정보 비식별조치 적용 지원

4. 과학데이터 품질 관리 체계 수립

과학데이터 활용을 위한 가장 중요한 요소는 데이터의 품질이다. 전문성이 높고 다양하고 복잡한 과학데이터의 특성상 데이터의 품질과 신뢰성은 해당 분야의 전문가가 아니면 검증의 한계가 있으며, 일괄적인 품질 관리는 현실적으로 어렵다. 따라서, FAIR (Findable, Accessible, Interoperable, Re-useable) 원칙[23]에 따른 데이터 관리를 통해 연구자들이 관련 과학데이터에 손쉽게 접근하고 이용할 수 있는 환경을 만들고 공개·활용을 통한 품질 검증 체계를 갖추는 것이 효과적인 품질 관리 방안이 될 수 있다.

또한 과학데이터는 연구의 객관성과 진실성을 검증하기 위한 기본 자료이며, 연구자 스스로 보존해야 할 중요한 자료이므로, 연구 산출물 중 하나인 과학데이터의 품질과 신뢰성에 대한 1차적인 책임은 연구자(기관)에 있음을 연구 윤리적 관점에서 강조할 필요가 있다.

국가 차원에서는 과학데이터 품질 관리를 위한 기준, 가이드라인 수립 및 품질 지원체계가 마련되어야 하며, 고품질의 데이터를 생산하고 연구자들이 손쉽게 활용할 수 있도록 지원하는 분야별 데이터센터의 역할 또한 강화해야 한다.

표 4. 주요 추진 과제 (4)

구분	주요 추진 내용
관리체계	· 과학데이터 공유 · 활용 전주기에 걸친 품질관리 및 검증 방안 수립 · 과학데이터 품질 관리 기준 및 가이드라인 제공 · 국가 과학데이터 메타데이터 표준 수립 및 상호운용성 확보
인프라	· 고품질 데이터 생산 및 활용 지원을 위한 분야별 데이터센터 구축 · 육성 · 과학데이터 품질 관리 지원 체계 마련

5. 국가 차원의 과학데이터 관리 · 공유 인프라 구축

과학데이터의 접근성 향상 및 공유 활성화를 위해 기반 인프라 제공이 필수적이다. 과학데이터 수집, 관리, 보존 및 공유를 지원하는 국가 차원의 과학데이터 플랫폼을 구축하고 국내외 데이터 및 서비스 연계를 통해 효과적인 과학데이터 공유·활용 환경 및 연구 협업 환경을 마련해야 한다.

또한 연구자가 과학데이터를 체계적으로 관리·공유할 수 있도록 지원하는 기준 및 가이드라인을 수립하고 수립된 가이드라인에 따라 과학데이터 관리·공유 플랫폼을 운영하고 활용을 지원할 조직체계 마련도 필요하다.

표 5. 주요 추진 과제 (5)

구분	주요 내용
법제도	· 국가 과학데이터 관리 · 공유 기반 구축을 위한 법적 근거 마련
관리체계	· 과학데이터 관리 · 공유 가이드라인 수립 · 과학데이터 관리 공유 지원 체계 마련 및 인력 양성
인프라	· 국가 과학데이터 수집 · 관리 · 보존 · 공유 플랫폼 구축 · 기관별 과학데이터 관리 · 활용 인프라 구축 · 국내외 과학데이터 연계 체계 마련 · 과학데이터 분석 · 활용 및 협업 환경 마련

6. 과학데이터 공유 · 활용 인식 제고

장기적 관점에서 과학데이터 공유·활용이 연구 문화로 정착될 수 있는 분위기 조성이 중요하다. 연구자들이 과학데이터 공유·활용 필요성을 인식하고, 자발적 참여를 유도할 수 있도록 성공사례를 발굴하여 홍보해야 한다. 또한, 전문 인력 양성 및 다양한 교육 프로그램 실시 등 과학데이터 공유·활용에 대한 연구자들의 인식 제고 방안을 지속적으로 추진해야 한다.

표 6. 주요 추진 과제 (6)

구분	주요 추진 내용
관리체계	· 과학데이터 공유 · 활용 성공사례 발굴 · 데이터 공유 인식 개선을 위한 교육 및 홍보 · 과학데이터 관리 · 공유 · 활용 관련 이해관계자들이 참여하는 오픈연구데이터포럼 운영 · 데이터 사이언스 전문 인력 양성

7. 역할과 책임

과학데이터의 공유·활용은 단계적으로 장기적 관점에서 접근해야 한다. 법제도적 기반과 동시에 국가 차원에서 갖추어야 할 인프라 구축이 먼저 선행되어야 한다. 분야별 데이터센터를 중심으로 데이터의 활용이 증진되고, 개별 연구자의 데이터 공유 문화로 확산될 수 있어야 한다.

국가 차원에서 과학데이터 공유·활용 체계가 작동하기 위해서는 데이터 생산자, 관리자, 연구자, 연구펀딩기관 등 다양한 이해관계자간의 협력이 필수적이다. 하지만, 이해관계자들의 과학데이터 관리 및 공유 경험이 거의 없고 심리적 장벽이 높으므로 제도 및 인프라 지원과 함께 충분한 논의와 의견수렴을 통한 점진적 접근이 필요하다. 정책 의사결정기구부터 연구수행기관까지 과학데이터 공유·활용 정책 수립 및 추진을 위한 역할을 정의하고, 각 역할을 수행할 기관간 협력체계 구축이 신중하게 접근되어야 한다. 또한 효율적이고 안정적인 정책 추진을 위해 현행 연구개발사업체계를 활용하고 정책 실행을 전문적으로 지원할 수 있는 전문지원기관 지정을 고려할 필요가 있다.

V. 결론

국가R&D과제에서 재사용 가능한 다양한 유형과 크기의 많은 과학데이터가 생산되고 있다. 그러나, 연구자 또는 연구실 수준에서 개별적으로 단순하게 저장·관리되고 있으며, 연구자 개인의 인적 네트워크를 통해 부분적으로 공유되고 있는 실정이다. 이는 데이터의 구축 자체가 성과로 인정받고 있지 못하며, 추가적인 데이터 관리, 품질 및 신뢰성 책임 이슈 등 데이터 관리·공유가 연구자에게 큰 부담으로 인식되고 있기 때문이다.

오픈 연구데이터는 새로운 과학적 발견을 촉진하고 연구 신뢰성 향상을 위한 세계적 흐름으로 자리매김하고 있다. 우리나라에서도 국가 차원의 과학데이터 관리·공유 체계 마련이 시급한 상황이다. 과학데이터의 개방·공유를 활성화하기 위해서는 과학데이터를 연구성과로 인정받을 수 있는 제도 개선이 필요하며, 연구자들이 보다 손쉽게 데이터를 관리·활용할 수 있도록 국가 차원의 데이터 관리·활용 지원체계를 마련해야 한다. 또한 국가R&D과제의 연구제안시 해외에서 이미 시행중인 데이터 관리 계획 제출을 우선적으로 시행할 필요가 있다.

다양한 이해관계자들의 과학데이터 공유·활용에 대한 입장도 분야, 문화, 전문성, 데이터 관리·공유를 바라보는 시각 등에서 많은 차이가 있는 상황이다. 따라서 오픈 연구데이터에 대한 공감대 형성을 위한 노력도 지속적으로 추진해야 한다.

향후 인문사회 분야를 포함하여 연구분야별 특성, 생산·활용현황, 공유 인식, 요구사항 등 세부적인 현황 파악 및 비교 분석을 위해 보다 많은 연구자를 대상으로 하는 현황 조사가 정기적으로 수행·발표될 필요가 있다. 또한 데이터 관리계획 실행 방안, 데이터 활용 규제 완화 방안 등 각종 정책, 가이드라인, 표준화 방안 등에 대한 연구, 데이터 기반 인프라 연구도 분야별 특성을 고려하여 다양한 관점에서 지속적으로 이루어져야 한다.

참 고 문 헌

[1] T. Hey, S. Tansley, and K. Tolle, *The Fourth Paradigm: Data-Intensive Scientific Discovery,* Microsoft Research, 2009.

[2] OECD, “Making Open Science a Reality,” OECD Science, Technology and Industry Policy Papers, No.25, 2015.

[3] The Royal Society, *Science as an open enterprise,* The Royal Society Science Policy Centre Report, 2012.

[4] J. Salmi, *Study on Open Science : Impact, Implications and Policy Options,* European Commission Report, 2015.

[5] 신은정, “오픈 사이언스(Open Science)에 관한 OECD 논의 동향과 시사점,” STEPI 동향과 이슈, 제22호, 2015.

[6] J. Holdren, *Increasing Access to the Results of Federally Funded Scientific Research,* White House OSTP Memorandum, 2013.

[7] Open Research Data Forum, *Concordat on Open Research Data,* Open Research Data Forum Report, 2016.

[8] 김지현, “국외 정부연구비지원기관의 연구데이터 관리정책 분석 - 미국, 영국, 캐나다, 호주를 중심으로,” 한국문헌정보학회지, 제47권, 제3호, pp.251-274, 2013.

[9] 윤종민, “과학데이터의 공유 및 활용 촉진을 위한 법적 과제,” 법학연구, 제27권, 제1호, pp.597-625, 2016.

[10] 심원식, “국가 차원의 연구데이터 관리체계 구축을 위한 로드맵 제안,” 한국문헌정보학회지, 제49권, 제4호, pp.355-378, 2015.

[11] N. Beagrie and J. Houghton, *The Value and Impact of the European Bioinformatics Institute,* EMBL-EBI Report, 2016.

[12] N. Beagrie and J. Houghton, *The Value and Impact of Data Sharing and Curation,* JISC Report, 2014.

[13] M. Baker, “Is There a Reproducibility Crisis?,” Nature, Vol.533, No.7604, pp.452-454, 2016.

[14] C. Begley and L. Ellis, "Raise standards for preclinical cancer research Clinical trials might be based on findings that cannot be replicated," Nature, Vol.483, No.7391, pp.531-533, 2012.

[15] L. Freedman, I. Cockburn, and T. Simcoe, "The Economics of Reproducibility in Preclinical Research," PLOS Biology, Vol.13, No.6, 2015.

[16] Open Science Collaboration, "Estimating the reproducibility of psychological science," Science, Vol.349, No.6251, 2015.

[17] Economist, "Excel errors and science papers," http://www.economist.com/node/21706466/, 2016.

[18] C. Tenopir, D. Hughes, and S. Allard, "Research Data Services in Academic Libraries: Data Intensive Roles for the Future?," Journal of eScience Librarianship, Vol.4, No.2, 2015.

[19] E. Barsky, "Research Data Management Survey : Science and Engineering," UBC Library Report, 2015.

[20] K. Shearer and F. Furtado, *COAR Survey of Research Data Management: Results,* COAR Report, 2017.

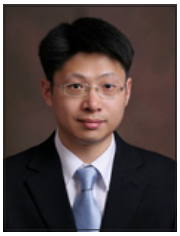
[21] 김문정, 김성희, "과학기술분야 연구자의 연구데이터 공유의 영향요인에 대한 연구," 한국문헌정보학회지, 제49권, 제2호, pp.313-334, 2015.

[22] 김지현, "데이터 관리와 공유에 대한 대학 연구자들의 인식에 관한 연구," 한국문헌정보학회지, 제49권, 제3호, pp.413-436, 2015.

[23] M. Wilkinson et al., "The FAIR Guiding Principles for scientific data management and stewardship," Scientific Data, Vol.3, 2016.

저 자 소 개

최 명 석(Myung-Seok Choi) 정회원



- 2005년 8월 : KAIST 전산학과(공학박사)
- 2005년 6월 ~ 현재 : 한국과학기술정보연구원 과학데이터연구센터 선임연구원

<관심분야> : 오픈사이언스, 연구데이터관리, 빅데이터분석 등

이 승 복(Seung-Bock Lee) 종신회원

- 1991년 2월 : 계명대학교 전산학과(공학사)
- 2008년 8월 : 충남대학교 전산학과(공학석사)
- 2014년 3월 : 충남대학교 전산학과(박사수료)

<관심분야> : 오픈사이언스, 인체정보, 과학데이터공동 활용 등

이 상 환(Sanghwan Lee) 정회원

- 2004년 8월 : 고려대학교 S/W공학과(공학석사)
- 1995년 4월 ~ 현재 : 한국과학기술정보연구원 과학데이터연구센터 책임연구원

<관심분야> : 과학데이터, 빅데이터, 오픈사이언스, 딥러닝 등