연구데이터 커먼즈 기반 연구 소프트웨어 재활용

송사광*+ 조민희*, 이미경*, 임형준*
*한국과학기술정보연구원 연구데이터커먼즈팀, +과학기술연합대학원대학교 응용AI학과
{esmallj, mini, jerryis, hjyim}@kisti.re.kr

Software Sharing & Utilization based on Research Data Commons

Sa-kwang Song*,+, Minhee Cho*, Mikyung Lee*, Hyung-Jun Yim*
Korea Institute of Science and Technology Information*, Univ. of Science and Technology+

요 약

오픈 데이터, 오픈 액세스, 오픈 소스 등 다양한 오픈 사이언스 운동이 과학기술계에서는 일반화되어 가고 있고, 최근에는 연구 소프트웨어의 재활용 체계 구축을 통한 R&D 산출물의 공유·활용 활성화에 관심이 높아지고 있다. 기존에는 연구자들이 생산한 연구 소프트웨어를 공개하고자 하여도 마땅히 공유할 시스템이나 체계가 없어, 개별 연구자 또는 부서 단위의 컴퓨터에 저장·관리하게 되어 시간이 지남에 따라 유실되는 등 공유·활용에 매우 취약하다. 해외 오픈사이언스 선진국의 경우는 오픈 사이언스 커먼즈(Open Science Commons), 연구데이터 커먼즈(Research Data Commons) 또는 오픈 사이언스 클라우드(Open Science Cloud) 등의 체계 구축을 통해 소프트웨어를 포함한 다양한 컴퓨팅 자원들을 공유·활용하고 있다. 이에 KISTI에서는 국가 연구데이터 커먼즈(KRDC: Korea Research Data Commons) 체계 구축을 통해, 이러한 컴퓨팅 자원의 공유·활용을 극대화하고자 하고 있다. 본 연구에서는 KRDC에 대한 소개와 KRDC 환경에서 소프트웨어 공유·활용 사례를 통해 이슈 및 향후 발전 방향을 소개한다.

Ⅰ. 서 론

최근 과학기술계에서 빈번히 회자되는 오픈사이언스는 R&D를 가속화시키는 중요한 운동으로, 오픈 액세스, 오픈 데이터, 오픈 협업, 오픈 소스 등 다양한 공유 활동이 포함된다[1]. 특히, 국가 R&D 과제와 같이 공적자금을 투입한 R&D 산출물은 개인의 자산이 아니라 국내 모든 연구자, 나아가 시민 과학자들에게 공유·활용되어야 할 공유 자산이라는 인식이 높아지고 있다. 특히, 연구개발의 핵심인 연구데이터와 이를 분석하는 소프트웨어 등은 국가 과학기술 경쟁력의 근간이 되는 중요 자산으로 정부 주도의 공유·활용 체계 구축이 필요하다[2-6]. 다행히 연구데이터의 수집·공유 체계는 2020년 오픈한 국가연구데이터플랫폼(DataON)[7]을 통해 다양한 분야의 연구데이터를 통합 관리 및 검색할 수 있도록 지원하고 있다. 그러나, 이러한 연구데이터를 분석활용하기 위한 연구 소프트웨어는 개별 연구자에게 관리가 일임되어 재활용되기 어렵고, 연구의 재현성 검증에도 거의 활용되지 않고 있다. 2021년 수행(KISTI)한 '연구데이터 활용 환경 개선' 설문(670명)[2]에 따르면, 컴퓨팅 리소스의 효율적인 공유·활용을 위한 체계 구축이 필요하다는 것에는 대부분(약 85%)의 연구자들이 동의하였다. 또한 분석도구나 컴퓨팅 자원이 공유된다면 활용하고자 하는 연구자가 약 92% 이상으로 나타날 정도로 연구데이터 높은 선호도를 볼 수 있었다. 즉, 많은 연구자들은 이미 연구 소프트웨어나 분석도구, 컴퓨팅 자원의 공유·활용을 적극적으로 원하고 있다고 할 수 있다. 이러한 필요에 대응하기 위해서 국내 컴퓨팅 자원의 용이한 수집, 공유, 활용을 위한 제도·정책의 마련과 통합 플랫폼 등의 개발이 필요하다.

본 논문에서는 컴퓨팅 자원의 공유·활용에 대한 필요성에 대응하기 위해 국가 연구데이터 커먼즈(KRDC: Korea Research Data Commons)를 소개하고 KRDC 구축을 위한 표준화된 미들웨어인 KRDC 프레임워크 및 이를 기반으로 개발된 활용 사례를 통해 이슈와 향후 발전 방향을 논한다.

Ⅱ. 국가 연구데이터 커먼즈(Korea Research Data Commons)

연구데이터 커먼즈(Research Data Commons)는 연구데이터 활용을 극대화하기 위해 상호 운용 가능한 고품질의 컴퓨팅 리소스에 원활한 사용성(usability)을 제공하는 신뢰할 수 있는 공통 활용 체계로 정의한다. 여기서 사용성이라 함은 FAIR(Findable, Accessable, Interoperable, Reusable)[8] 원칙, 연구과정에서 생산된 연구데이터 및 디지털 개체에 대한 신뢰가능한 재사용성 등을 의미한다. 이는 간략히 표현하면, 연구데이터와 연구데이터 관련 컴퓨팅 리소스의 연합·활용 체계로 이해할 수 있다.[2,3]

연구데이터 커먼즈(Research Data Commons)는 2018년 국가과학기술심의회 승인을 받아 추진되고 있는 "혁신성장 촉진을 위한「연구데이터 공유·활용 전략」"을 기반으로 하고 있다. 지금까지 주로 연구데이터의 수집·공유에 초점을 두고 국가연구데이터플랫폼인 DataON의 구축, 소재·바이오 등의 분야별 연구데이터 플랫폼 구축, 국가출연연구소 등의 연구데이터 플랫폼 구축 등에 집중해 왔다. 그러나 연구의 핵심인 연구데이터는 효율적으로 활용되어야만 그 가치가 배가되고 혁신을 만들어낼 수 있기에, 이제부터는 데이터를 분석·처리하는 연구 소프트웨어의 공유에 관심과 노력을 기울여, 축적되어 가는 연구데이터를 더 효율적으로 활용할 수 있는 체계 구축에 집중해야할 때이다.

그리하여, KISTI에서는 2021년부터 국가 연구데이터 커먼즈(KRDC: Korea Research Data Commons) 체계를 구축해 오고 있고, 그 일환으로 KRDC 프레임워크라는 연구 소프트웨어 공유·활용 지원을 위한 미들웨어를 개발하였다. 이는 클러스터 운영체제로 알려진 쿠버네티스(Kubernetes)를 기반으로, 연구 소프트웨어를 앱 형식으로 포장하여 다중 클러스터 환경에서 다수의 사용자가 손쉽게 재활용할 수 있는 환경을 지원하기 위한 프레임워크이다. 다양한 연구 소프트웨어를 워크플로우로 엮

어서 특정 작업의 수행을 용이하게 하기 위한 소프트웨어 시스템이라 할 수 있다. 워크플로우라 함은 하나 이상의 연구 소프트웨어 앱을 조합하여 사용자가 원하는 작업을 수행하기 위한 작업 흐름을 정의한 그래프를 말한다. KRDC 프레임워크는 대부분 오픈 소스를 기반으로 구축되어 향후 확장성 및 이식성 등을 극대화하고자 하였다.

Ⅲ. KRDC 프레임워크 활용

KRDC 개념을 구현하고, 다중 클러스터 환경에서 연구 소프트웨어의 공유·활용을 촉진하기 위한 구현체가 KRDC 프레임워크이다. KRDC 프레임워크에서는 연구에서 많이 사용되는 도커 컨테이너(Docker Container) 환경에 연구 소프트웨어를 설치하도록 가이드 한다. 즉, 연구 소프트웨어를 개발하는 연구자들이 자신만의 환경에 적합하도록 도커 환경을 구축하고, 그 위에 자신의 연구 소프트웨어를 개발하도록 하여 많은 연구자들이 쉽게 개발 및 활용할 수 있는 공통적인 연구 소프트웨어 개발환경을 제공한다. 대부분의 연구 소프트웨어가 입력 및 출력, 기타 파라미터를 받아서 전처리, 분석, 가시화, 인공지능 등의 태스크를 수행한다는 것을 고려할 때 입출력 정보만을 공유하여도 재활용이 가능하다. 이런 이유로 입출력 정보를 기반으로 앱들을 연계하여 작업을 수행하는 워크플로우 엔진들이 다양한 곳에서 활용되고 있고, DataON에서도 이미 연구 소프트웨어를 앱으로 구현하여 웹 GUI 기반 워크플로우 분석 환경을 운영하고 있다. 다만 DataON과 비교할 때, KRDC 프레임워크의 가장 큰 차이점은 복수의 클러스터 환경에서 연구 소프트웨어를 조합하여 작업을 수행할 수 있다는 점이다.

그림 1은 이러한 연구 소프트웨어가 쿠버네티스 환경에서 도커 컨테이너에 탑재되는 경우를 단순화한 개념도이다. 일반적으로 연구 소프트웨어는 입력 데이터 및 입력 파라미터를 받아 특정 작업을 수행하고 출력으로 데이터를 생산하거나 값을 출력하는 등의 작업을 한다.

그림 1 쿠버네티스의 컨테이너에 탑재된 연구 소프트웨어 개념도

이러한 입출력 패턴만 숙지한다면 하나 이상의 연구 소프트웨어를 연결하여 자신만의 워크플로우를 구성하여 여러 작업을 순차적 또는 병렬적으로 수행할 수 있다. 쿠버네티스 생태계에서는 이를 위해 Argo[9]라는 쿠버네티스 워크플로우 엔진을 제공하고 있고, KRDC 프레임워크에서도 Argo를 사용하고 있다. 아래는 실제 Argo 상에, BCI(Brain Computer Interaction) 분야의 전처리 및 딥러닝 연구 소프트웨어를 탑재하여 동작시킨 화면이다.

본 프레임워크를 사용하여 연구 소프트웨어를 공유하면 재활성을 높일 수 있을 뿐만 아니라, 연구의 재현성 확보에도 큰 도움이 된다. 특히, 연구 소프트웨어를 앱 형태로 등록함으로써, 입출력만 바꿔가며 다양한 조합의 워크플로우를 만들어 손쉽게 연구를 수행할 수 있을 뿐만 아니라, 기존 워크플로우에 자신만의 모듈(앱)을 추가 개발하여 확장성도 높일 수 있다. 이러한 체계가 장점만 있는 것은 아니다. 자신만을 위한 연구 소프트웨어를 개발하는 것보다, 당연히 연구소프트웨어를 공유 앱으로 만들 때 더욱 많은 수고가 드는 것은 피할 수 없다. 또한 연구자나 소속 기관들의 연구 소프트웨어에 대한 공개 거부감은 또 다른 차원의 문제라고 할 수 있다. 이는 제도적 정책적 지원과 함께 오픈 사이언스 문화의 성숙을 기다릴 수 밖에 없다.

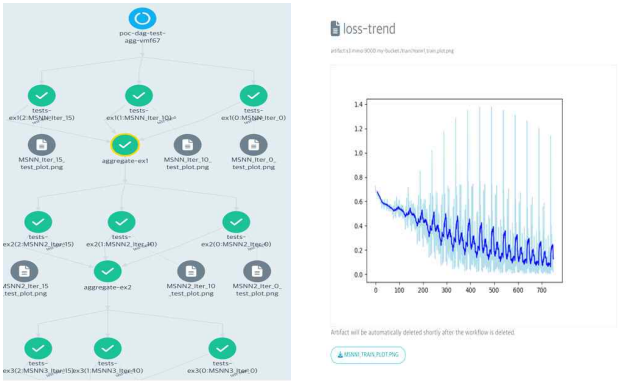

그림 2 KRDC 프레임워크에 탑재된 BCI 연구의 워크플로우 및 결과

Ⅲ. 결론

본 논문에서는 KRDC 개념과 KRDC 프레임워크를 소개하고, KRDC 프레임워크에 동작하는 사례를 통해 연구 소프트웨어의 공유·활용에 대한 방안을 제시하였다. 연구데이터의 공유를 넘어 연구 소프트웨어의 공유가 활성화되기까지는 꽤 많은 시간이 필요 하겠지만 국내외의 오픈 사이언스 트렌드가 과학기술 발전에 필연적인 움직임이라는 것을 고려할 때, 국내의 소프트웨어 공유·활용 체계 구축 및 활성화가 근 시일내에 정착할 것이라 기대한다.

ACKNOWLEDGMENT

본 연구는 한국과학기술정보연구원 연구데이터와 인프라의 공유·활용 체제 구축(K-22-L01-C03-S01) 사업의 지원을 받아 수행된 연구임.

참 고 문 헌

[1] Woelfle, M.; Olliaro, P.; Todd, M. H. (2011). "Open science is a research accelerator". Nature Chemistry. 3 (10): 745 - 748. doi:10.1038/nchem.1149

[2] Sa-kwang Song, Dongmin Seo, "Revitalization Plan for R&D Research Outcomes: focusing on research data", Korea International Conference on Convergence Content, Jeju Special Self-Governing Province, Korea, 2021

[3] 송사광, 서동민, "연구데이터 공유활용 제도 및 정책 현황", GeoAI데이터학회 추계학술대회, 부산, pp. 118-119, 2021

[4] 송사광, 조민희, 이미경, 임형준(2022). "연구데이터 활용성 극대화 위한 컴퓨팅 리소스 공유활용 체계", 한국정보통신학회 추계학술대회, 제주

[5] 조민희, 이미경, 임형준, 송사광(2022). "국가연구데이터커먼즈 서비스를 위한 데이터모델 연구", 한국정보통신학회 추계학술대회, 제주

[6] 이미경, 조민희, 송사광, 임형준(2022). "컴퓨팅 리소스 관리를 위한 표준 메타데이터 스키마 설계", 한국정보통신학회 추계학술대회, 제주

[7] 국가연구데이터플랫폼 DataON: Available : http://www.dataon.kr

[8] FAIR Principles: Available : https://www.go-fair.org/fair-principles/

[9] Argo Workflows - The workflow engine for Kubernetes: Available https://argoproj.github.iohttps://ardc.edu.au/