국가연구데이터커먼즈 서비스를 위한 데이터모델 연구

조민희[1*] · 이미경[1] · 송사광[1,2] · 임형준[1]

[1]한국과학기술정보연구원 연구데이터공유센터 · [2]과학기술연합대학원대학교 응용AI학과

Data Model Study for National Research Data Commons Service

Minhee Cho[1*] · Mikyoung Lee[1] · Sa-kwang Song[1,2] ·Hyung-Jun Yim[1]

[1]Research Data Sharing Center, KISTI · [2]Dept. of Applied AI, UST

E-mail : {mini, jerryis, esmallj, hjyim}@kisti.re.kr

요 약

국가연구데이터커먼즈는 연구데이터 활용 극대화를 위해 컴퓨팅 인프라, 데이터 분석을 위해 사용되는 SW, Toolkit, API, 서비스 등과 같은 분석리소스를 연구데이터와 함께 배치하여 연합 활용될 수 있는 체계를 구축하는 것을 목표로 한다. R&D 과정에서 연구 출판물, 연구데이터에 대한 공유·활용 체계는 이미 잘 알려져 있다. 하지만 데이터와 밀결합된 소프트웨어, 컴퓨팅 인프라들을 공유하고, 활용할 수 있는 환경은 미미하고, 관리체계가 없는 실정이다. 본 연구에서는 데이터 중심의 R&D 연구 과정에서 필요한 디지털 연구 자원 정보를 체계적으로 관리하기 위하여 데이터모델을 설계한다. 이는 국가연구데이터커먼즈 서비스에서 디지털 연구 자원 정보들을 등록하고 관리하는데 활용될 예정이다.

ABSTRACT

National Research Data Commons aims to build a system that can be used jointly by arranging analysis resources such as computing infrastructure, software, toolkit, API, and services used for data analysis together with research data to maximize the use of research data. do. The sharing and utilization system for publications and research data in the R&D process is well known. However, the environment in which data and tightly coupled software and computing infrastructure can be shared and utilized is insignificant and there is no management system. In this study, a data model is designed to systematically manage information on digital research resources required in the data-oriented R&D research process. This will be used to register and manage digital research resource information in the National Research Data Commons Service.

키워드

Research Data Commons, Data Model, Open Science, Interoperability, FAIR, Digital Resource

Ⅰ. 서 론

최근 발표된 유네스크 ‘오픈 사이언스에 대한 권고안’에서는 오픈사이스는 디지털 기술 발전에 따라 연구의 성과와 과정을 보다 폭넓고 개방적으로 공개·공유하려는 지향과 실천까지 포함하고 있다[1]. 펜데믹 상황을 겪으면서 국가적으로 R&D 혁신을 위한 분야간, 지역간, 국가간 장벽을 무너뜨리고, 연구 과정에서 활용되고 생산되는 다양한 종류의 연구 결과물들이 공유될 수 있는 연구 디지털 환경에 대한 수요가 높아지고 있다. R&D 연구 결과물로서 논문, 연구데이터에 추가적으로 컴퓨팅 인프라와 데이터 분석에 사용되는 SW, Toolkit, API, 서비스 등과 같은 분석리소스까지 관리할 수 있는 체계가 필요하다. 또한 이들 간의 상호 운용이 가능하도록 표준 및 정책을 마련하여 연결을 강화하는 체계를 마련해야 할 것이다.

이러한 오픈사이언스를 지원하기 위하여 KISTI에서는 ‘국가연구데이터커먼즈 기반의 공유, 활성화’를 위한 프로젝트를 수행하고 있다. 연구데이터뿐만 아니라 활용을 위한 분석인프라 서비스 정보를 함께 제공함으로써 연구 전주기적 과정에 전사적으로 연구자를 지원하기 위함이다. 분석인프라 서비스는 연구 전과정에서 필요한 서비스이므로

* corresponding author

본 연구에서는 연구리소스라고 정의한다.

이러한 연구리소스 정보는 온라인에서 접근 가능하고 활용가능한 서비스이다. 물리적 형태의 서버, 스토리지, 네트워크 등의 ID 인프라를 제공하는 서비스, 다양한 애플리케이션을 개발할 수 있는 개발환경 서비스, 온라인 분석 애플리케이션 서비스 등이 다. 이 가운데 온라인 분석 애플리케이션 서비스는 연구자들이 만든 연구소프트웨어가 온라인에서 하나의 서비스 형태로 운영되는 것이다.

연구소프트웨어는 R&D 과학적 산출물로서의 개념에 초점을 맞출 수 있다. 연구 논문에 게시되거나, 배포된 결과를 생성하기 위해 구축되고 사용된 소프트웨어로 정의할 수 있다. 소스코드, 컴파일 코드 등이 포함된 파일세트, 문서, 명세서, 테스트집합으로 구성될 수 있다[2,3].

본 연구에서는 기존에 관리되지 않았던 새로운 디지털 오브젝트인 연구리소스를 정의하여 확장하려고 한다[4]. 기존의 DataON 플랫폼에 정의된 연구데이터, 논문, 연구소프트웨어와 같은 연구결과물을 컬렉션으로 그룹화하여 정의하고 이를 연구자들이 직접적으로 온라인에서 접근/활용하는 서비스 객체로 '리소스'오브젝트를 정의하여 이들간의 의미관계를 정의한다. 국가연데이터커먼즈 서비스를 위한 데이터 모델을 3장에서 제안한다.

II. 관련연구

유럽의 오픈사이언스 및 오픈 이노베이션을 지원하기 위해 2015년 유럽연합집행위원회(EC)가 클라우드를 기반으로 연구데이터 저장, 관리, 분석, 재사용을 가능하게 하는 가상환경을 제공하는 서비스를 구현하기 위한 EOSC(Europe Open Science Cloud) 프로젝트를 진행 중에 있다. EOSC 포털은 유럽의 모든 과학정보 및 리소스를 접근하기 위한 게이트웨이로서의 진입점 역할을 한다. EOSC '서비스 제공자' 및 '사용자'들을 위해 제공되는 카탈로그 서비스에서는 그림1의 정보 모델로 관리되고 있는 메타 정보들이 열람 가능하다[4].

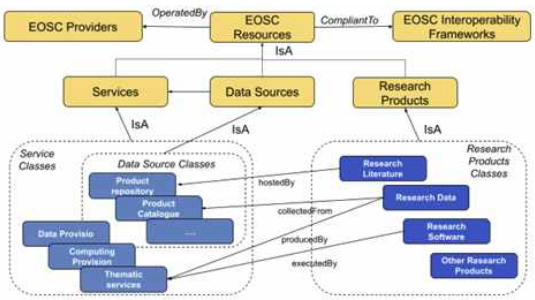

그림 1. EOSC 정보 모델[5]

EOSC 포털을 통해 사용가능한 디지털 자산을 리소스로 정의하고 있다. 리소는 연구산출물(논문, 연구데이터, 연구소프트웨어, 기타생산물), 데이터 소스, 서비스 등에 해당한다. 리소스(Resources)와 EOSC 제공자(Providers)는 다양한 정보 블록으로 데이터를 관리한다. 리소슬 생산하는 제공자는 기본정보, 마케팅정보, 과학적 분류정보, 위치정보, 연락처정보, 협력정보로 구성되어 있고, 리소스는 기본정보, 마케팅정보, 분류정보, 언어 정보, 국가정보, 연락처정보, 성숙도 정보, 협력정보, 속성정보, 펀딩정보, 관리정보, 접근/주문정보, 가격정보 등으로 구성되어 있다. 리소스 활용도를 높이기 위한 제공자들의 마켓정보와 제품으로서의 성숙도 및 가격정보들을 포함하는 것이 특징이다.

III. 국가연구데이터커먼즈 데이터 모델

앞에서 설명한 KRDC(Korea Reseach Data Commons) 개념을 기반으로 KRDC-Hub는 데이터 중심의 과학적 연구 수행에 필요한 데이터 및 분석 인프라(예: 스토리지, 컴퓨팅, 모델, 소프트웨어, 출판물)에 접근 가능한 모든 디지털 오브젝트에 대한 정보를 모으는 허브로서의 역할을 수행하고, 원활한 접근성 및 사용성을 제공하는 포털이다.

KRDC-Hub를 통해 제공되는 카탈로그 서비스를 통해 제공되는 데이터 구축을 위하여 디지털 연구자원 정보들을 다음과 같이 정의한다. 우리가 관리해야할 오브젝트는 Provider(기관, 사람), Project(과제), Collection(연구성과물), Resource(연구리소스), Policy(정책) 등이다.

- Provider: 컬렉션을 생산하거나 서비스를 개발/운영하는 연구자 혹은 조직 정보
- Collection: 연구데이터, 연구출판물, 연구소프트웨어 등 연구과정에서 산출되는 성과물 정보
- Project: 국가 R&D 연구 과제 정보
- Resource: 컬렉션을 생성/관리하는데 필요한 디지털 연구 자원에 대한 정보
- Policy: Resource 활용을 위한 가격, 라이센스, 권한 등의 정책 정보

이렇게 정의된 오브젝트들은 리소스의 PID정책에 따라 고유의 ID로 관리될 것이다. 그림 2와 같이 각 데이터들은 특성 정보를 포함하고, 오브젝트들간의 관계를 통해 연관된 정보들을 파악할 수 있다. 데이터모델을 활용하여 모든 데이터의 관계를 따라가다 보면, 데이터와 관련된 히스토리를 쉽게 파악할 수가 있다.

예를 들면 아래와 같은 연구과정은 데이터모델로부터 쉽게 해석할 수 있다.

> 연구자가 생산한 'A'논문은 'B' 데이터셋과 'C' 연구소프트웨어를 활용하여 생산이 되었고, 'D'기관이 'F'프로젝트를 통해 수행된 연구결과이다. 이러한 컬렉션들의 데이터는 'G','H','I'리소스 서비스에서 관리되고 있다.

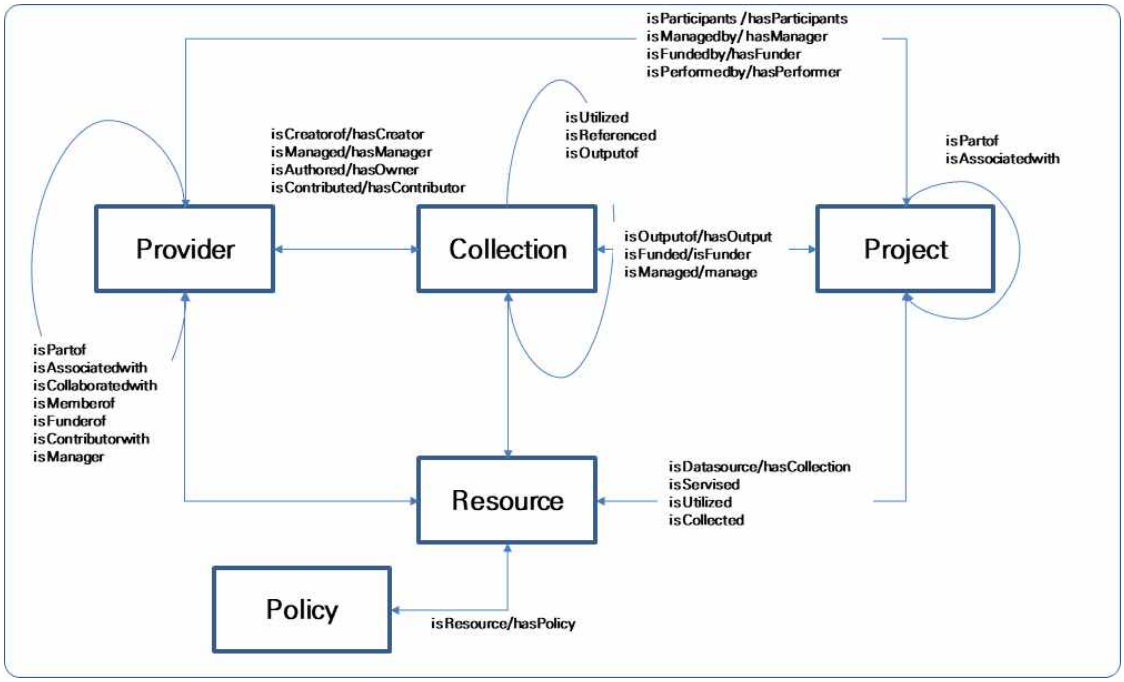

그림 2. KRDC 데이터 모델

Ⅳ. 결 론

본 논문에서는 R&D 연구 과정에서 필요한 디지털 연구 리소스들을 데이터 모델로 표현하고, 이를 활용하기 위한 방안을 제시했다. 아직 공공 R&D 연구결과물로 리소스 및 컬렉션의 연구소프트웨어 등과 같은 데이터 관리 체계가 만들어지지 않아 데이터 수집 및 통합이 어려운 실정이다. KISTI는 우선 출연연, 전문센터를 중심으로 현재 관리되고 있지 않는 연구소프트웨어, 연구리소스 데이터들을 관리할 수 있는 체계를 준비하고 있다.

향후 체계가 완성된다고 하면, 모든 국내 디지털 연구 정보들이 한곳에 모여 국가연구데이터커먼즈 서비스에서 제공될 수 있을 것이다.

Acknowledgement

본 연구는 2022년도 한국과학기술정보연구원(KISTI) 주요사업 “연구데이터와 인프라의 공유·활용 체제 구축”과제로 수행한 것입니다.

References

[1] UNESCO Recommendation on Open Science [Internet]. Available : https://unesdoc.unesco.org/ark:/48223/pf0000379949.locale=en

[2] Gomez-Diaz T and Recio T. “Research Software vs. Research Data I: Towards a Research Data definition in the Open Science context,” *F1000Research*, 2022.

[3] Towards FAIR principles for research software [Internet]. Available : https://content.iospress.com/articles/data-science/ds190026.

[4] M. H. Cho , S. K. Song, H. J. Yim, “Exploring National Science and Technology using Research Resource Knowledge Graph,” *Proceedings of the Korean Institute of Information and Commucation Sciences Conference*, p. 621-623, Oct. 2021.

[5] EOSC Research Product Profile [Internet]. Available : https://eosc-portal.eu/providers-documentation/eosc-research-product-profile