

# 연구리소스 지식그래프를 활용한 국가과학기술정보 탐색

조민희\* · 임형준 · 송사광

한국과학기술정보연구원

## Exploring National Science and Technology using Research Resource Knowledge Graph

Minhee Cho\* · Hyung-Jun Yim · Sa-kwang Song

KISTI

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

### 요 약

공공의 자금을 통해 산출된 연구 성과물을 온라인으로 공개하여 공유, 활용하는 오픈 사이언스 정책이 확산되고 있다. 오픈사이언스를 활성화기 위한 정책으로 공개되는 성과물을 쉽게 검색 및 액세스하고 재사용할 수 있도록 하는 연구 지원 서비스에 관심이 증대되고 있다. 연구자에게 풍부한 정보를 제공하기 위한 서비스 지원을 위하여 흩어진 다양한 성과물 데이터들 사이의 관계를 의미적으로 표현하기 위한 연구리소스 지식그래프 모델을 제안한다. 본 논문에서는 지식 그래프 구축을 통해 국가 R&D 과제정보, 연구자정보, 성과정보, 연구데이터정보 등을 의미적으로 연결해 연구개발에 필요한 전반적인 연구리소스 정보 탐색이 가능하도록 하여 국가연구데이터플랫폼 DataON의 서비스 향상에 기여한다.

### ABSTRACT

Open science policies are spreading that disclose, share, and utilize research results produced through government public funds. As a policy to revitalize open science, interest in research support services that allow easy search, access, and reuse of results is increasing. To support services to provide researchers with various information, we propose a research resource knowledge graph model to meaningfully express the relationship between the scattered various outcome data. In this paper, it contributes to the improvement of the service of the national research data platform DataON by meaningfully connecting national R&D task information, researcher information, performance information, and research data information.

### 키워드

연구데이터, 연구리소스, 지식그래프, 국가과학기술정보

### 1. 서 론

오픈사이언스는 모든 종류의 과학적 지식을 개방하고 공유하는 것을 말한다. 이러한 개방을 추진하기 위하여 연구수행과정에서 요구되거나 산출되는 과제정보, 성과정보, 연구데이터, 협력 기관 및 연구자 등의 정보를 통합 저장, 관리되거나, 찾고자 하는 정보들의 관련 정보를 한눈에 파악할 수 있는 시스템이 필요하다. 이를 통해 연구자들의 연

구기획, 산출물 활용 등에 시간을 단축할 수 있을 것이다.

또한 유사 연구를 진행하는 기관, 연구자들을 한눈에 파악할 수 있다면, 이전에는 어려웠던 융합 연구 협력 파트너도 쉽게 발견할 수 있어 새로운 융합 연구를 실행할 기회가 확대될 것이다.

본 연구에서는 과제정보, 연구자정보, 성과정보, 연구데이터정보 등 과학기술정보를 기반으로 정보들 간의 관계를 표현하는 연구리소스 지식그래프를 제안한다.

\* corresponding author

## II. 연구리소스 지식그래프

연구자들은 국가 R&D 연구수행을 통해 논문, 소프트웨어, 연구데이터 등의 성과물을 산출하고 과제 단위로 매년 성과정보를 등록하고 있다. KISTI는 R&D과정에서 산출된 이러한 성과물정보를 NTIS\*, ScienceON\*\*, DataON\*\*\*등을 통해 서비스하고 있다. 그 중 국가연구데이터플랫폼(DataON)은 연구데이터를 저장/관리하고 활용하기 위한 다양한 서비스를 제공하고 있다. 현재 DataON의 국내 연구데이터 서비스 건수는 3,400여건이다. 연구데이터 공유에 대한 인식이 아직 정착되지 않아 기탁건수는 미미한 실정이다. 하지만 연구데이터의 중요성이 강조되고 있고, 연구데이터 공유 문화가 확산되면서 데이터의 양이 증가하고 있는 추세에 있다.

연구리소스 지식그래프는 국내 연구데이터를 중심으로 관련된 R&D 과제, 성과(논문, SW 등), 연구자, 기관 등 정보들 간의 관계를 유기적으로 연결해놓은 것을 의미한다. 연구데이터를 중심으로 탐색을 통해 관련된 논문, 분석SW, 연구자, 기관, 유발과제 정보 등 관련정보를 유연하게 확장 가능하고 쉽게 접근 가능하다. 그래프를 구성하는 각 노드들은 데이터를 표현하는 다양한 속성정보를 포함하는 메타데이터 정보를 가지고 있다. 각 노드 정보의 데이터는 서로 다른 서비스에서 관리되고 있는 대상데이터로, 각각의 데이터를 한곳에 모아서 데이터들 간의 의미관계를 정의하고자 한다. 하지만 데이터들이 동일한 속성을 사용하는 정보가 빈약하여, 데이터들간의 의미관계를 파악하는 것이 쉽지 않다. 예를 들면, 동일한 연구자가 논문, 소프트웨어, 연구데이터를 생산하더라도, 이름이 같은 것만으로 동일 연구자임을 판단하는 것은 위험요소가 많다.

해외에서는 Research Graph 이니셔티브를 식별정보를 기반으로 그래프 모델 연구들이 추진되고 있다[1,2]. 연구데이터그래프 모델을 개발하기 위하여 연구자 식별 정보를 중심으로, 논문, 연구데이터, 연구자 등을 쉽게 연결하고 있다. 개별 연구자를 구별하는 글로벌 연구자 식별정보인 ORCID(Open Researcher and Contributor ID), 연구 성과물 식별정보인 PID(Persistent Identifier, PID)가 확대되어 활용되고 있고, ORCID, PID를 기반으로 개별 연구자의 연구 활동 및 성과물 정보가 연계되어 관리되고 있다.

DataON은 국가 R&D 예산을 통해 생산된 3천여건의 연구데이터가 저장 관리되고 있다. 연구데이터가 등록될 때 유발된 국가과학기술지식정보서비스(NTIS)의 과제정보, 과학기술지식인프라서비스

(ScienceON)의 논문정보를 포함하고 있다.

또한 DataON은 다양한 연구자 식별정보를 등록 받고 있다. 회원 가입시 연구자 식별정보인 NTIS 과학기술인등록번호, ORCID정보 등 다양한 식별정보를 입력 받고 있다. 연구자 정보를 중심으로 이미 보존 관리되고 있는 과제정보, 국내외 논문정보 등을 쉽게 연결하기 위함이다.

연계된 시스템의 모든 데이터 정보를 저장 관리하는 것이 아니라, 연계를 위한 최소한의 고유식별번호만 저장, 관리하여 정보를 확장한다.

시작단계에 있는 DataON은 연구데이터, 과제, 성과(논문,SW), 인력 정보를 연구식별자를 기반으로 연결할 수 있도록 시스템 기반을 갖추고 있다. 아직은 데이터 건수가 작아 연결이 미미하지만, DataON 활용이 활성화되어 연구데이터 등록 및 활용 사용자가 확대된다면 국내 연구자의 과학기술인등록번호를 중심으로 NTIS에 기구축된 과학기술 정보들과 연계하여 자동으로 확장이 가능해진다. 또한 해외 연구자 식별정보를 기반으로 해외 연구데이터, 논문 정보 등으로도 유연하게 확장이 가능해진다.

연구리소스 지식그래프 구축을 위하여 특성이 다른 데이터들 사이의 의미 관계를 표현할 수 있도록 데이터 노드, 관계를 정의한다. 그래프를 구성하는 노드 유형은 연구데이터(DataSet), 연구자(Researcher), 조직(Organization), 과제(Project), 논문 정보(Article), SW정보(Software) 등이 있다. 노드들은 출처로부터의 id, 내부 관리 id, pid 등 다양한 id를 관리할 수 있는 체계를 갖추고 있다. 각 노드 사이의 의미관계는 표 1과 같이 정의한다.

연구리소스 지식그래프 시각화를 위하여 오픈소스인 neo4j 라이브러리를 활용하여 그림2와 같이 표현 하였다. 데이터셋을 중심으로 유발 과제정보, 참여인력정보, 성과 논문 정보 등을 확인할 수 있다. 개체를 클릭하면, 개체와 관련된 데이터셋, 논문, 참여한 과제정보를 확장해서 계속적으로 확인할 수 있다. 지식그래프는 모든 데이터의 관계를 표현하고 있어 관계를 따라가다 보면, 데이터와 관련된 히스토리를 쉽게 파악할 수가 있다. 예를 들면 ‘A라는 연구데이터는 KISTI 연구소의 Dr.Smith 연구자에 의해 작성이 되었고, B라는 논문 성과를 출판하는데 A 연구데이터를 활용하였다. 해당 데이터와 논문은 C과제를 통해 산출되었다.’등의 정보를 해석할 수 있다.

표 1. 관계 정의 예

related To	SourceNode	TargetNode
IsAuthoredBy	DataSet	Researcher
IsOutputOf	DataSet	Publication
IsTransformed	DataSet	DataSet
IsLaterVersionof	DataSet	DataSet
IsOutputOf	DataSet	Project
Utilize	DataSet	Software

\* <https://ntis.re.kr>, 국가과학기술지식정보서비스  
 \*\* <https://scienceon.kisti.re.kr>, 학술정보서비스  
 \*\*\* <https://dataon.kisti.re.kr> 연구데이터서비스

HasParticipants	Project	Researcher
IsManagedBy	Project	Researcher
FunedBy	Project	Organization
PerformedBy	Project	Organization
IsOutputOf	Publication	Project
IsAuthoredBy	Publication	Researcher
IsParticipant	Researcher	Organization
CollectedFrom	DataSet Software	DataSource (데이터제공처, 저장소)

References

- [1] Aryani, A., Poblet, M., Unsworth, K., Wang, J., Evans, B., Devaraju, A., et al. "A Research Graph dataset for connecting research data repositories using RD-Switchboard", *Sci. Data*, 2018.
- [2] Manghi, Paol, Bardi, Alessia, The OpenAIRE Research Graph - *Opportunities and challenges for science*, 2019.

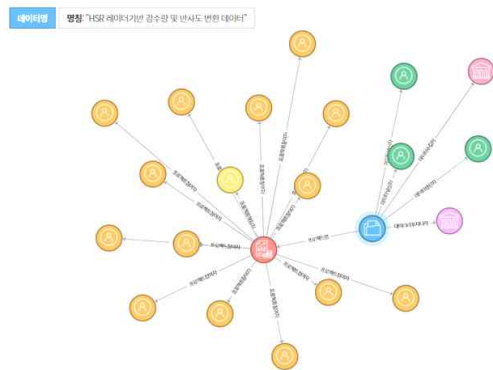


그림 1. 연구데이터 그래프 예

이미 많은 연구자 및 안정된 플랫폼 서비스를 하고 있는 NTIS, ScienceON 등의 서비스 데이터를 기반으로 신규 등록되는 DataON 연구데이터들이 연계되어 데이터들간의 관계가 정의된다면, 한눈에 국가과학기술정보를 확인할 수 있는 거대한 지식 맵이 될 것이고, 정책입안자, 연구자들의 연구생산성 향상에 기여할 것이다.

III. 결 론

본 논문에서는 국가과학기술데이터를 연계하여 통합 모니터링하기 위하여 연구데이터 그래프를 제안했다. 국가 R&D 예산을 통해 발주된 과제 산출물인 데이터셋을 중심으로 연구자, 조직, 성과 정보를 연계했다. 향후 자동화 방안을 모델링하여 연구데이터 그래프를 확대할 예정이다.

Acknowledgement

이 논문은 한국과학기술정보연구원(KISTI) 주요 사업 “연구데이터와 컴퓨팅 리소스의 공유·활용 체제 구축” 지원에 의함