JISTaP http://www.jistap.org
Journal of Information Science Theory and Practice

RESEARCH PAPER
J Inf Sci Theory Pract 10(Special): 25-38, 2022

eISSN : 2287-4577 pISSN : 2287-9099

https://doi.org/10.1633/JISTaP.2022.10.S.3

Development of a National Research Data Platform for Sharing and Utilizing Research Data

Youngho Shin
Korea Institute of Science and Technology Information (KISTI), Daejeon, Korea
E-mail: shinyh@kisti.re.kr

Jungho Um
Korea Institute of Science and Technology Information (KISTI), Daejeon, Korea
E-mail: jhum@kisti.re.kr

Dongmin Seo
Korea Institute of Science and Technology Information (KISTI), Daejeon, Korea
University of Science and Technology, Daejeon, Korea
E-mail: dmseo@kisti.re.kr

Sungho Shin*
Korea Institute of Science and Technology Information (KISTI), Daejeon, Korea
E-mail: maximus74@kisti.re.kr

ABSTRACT

Research data means data used or created in the course of research or experiments. Research data is very important for validation of research conducted and for use in future research and projects. Recently, convergence research between various fields and international cooperation has been continuously done due to the explosive increase of research data and the increase in the complexity of science and technology. Developed countries are actively promoting open science policies that share research results and processes to create new knowledge and values through convergence research. Communities to promote the sharing and utilization of research data such as RDA (Research Data Alliance) and COAR (Confederation of Open Access Repositories) are active, and various platforms for managing and sharing research data are being developed and used. OpenAIRE (Open Access Infrastructure for Research In Europe), a research data platform in Europe, ARDC (Australian Research Data Commons) in Australia, and IRDB (Institutional Repositories DataBase) in Japan provide research data or research data related services. Korea has been establishing and implementing a research data sharing and utilization strategy to promote the sharing and utilization of research data at the national level, led by the central government. Based on this strategy, KISTI has been building a Korean research data platform (DataON) since 2018, and has been providing research data sharing and utilization services to users since January 2020. This paper reviews the characteristics of DataON and how it is used for research by showing its applications.

Keywords: research data, research data platform, DataON, open data, open science, data platform

Received: April 25, 2022 **Revised:** May 11, 2022
Accepted: May 17, 2022 **Published:** June 20, 2022

***Corresponding Author:** Sungho Shin
https://orcid.org/0000-0001-9448-2589
E-mail: maximus74@kisti.re.kr

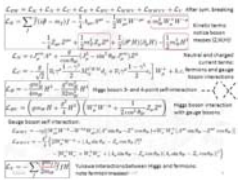
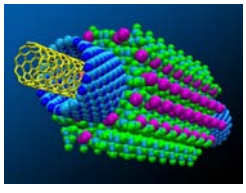
1. INTRODUCTION

Major developed countries have defined research data as follows. The OECD defines research data as factual data such as figures, texts, images, and sounds, and it is mainly used as a major source of scientific research and is essential to verify research results among scientific scholars (Pilat & Fukasaku, 2007). The US National Information Standards Organization (NISO) has defined it as data that occurs throughout the entire research cycle. The UK Joint Information Systems Committee (JISC) defines it as data generated in the entire research cycle in terms of data management. The Australian Research Data Commons (ARDC) defines it as data generated in the form of facts, observations, images, computer program results, records, measurements, or experiences. In the Republic of Korea's national R&D information processing standards, it is defined as data essential for verification of research results as factual data calculated through various experiments, observations, investigations, and analysis conducted in the course of carrying out R&D tasks. In other words, although the definition of research data by country is slightly different, the data generated during the entire research cycle is defined as research data.

With the Fourth Industrial Revolution, the use of data is emerging as the core of scientific and technological competitiveness. For example, Gartner (https://gartner.com/en/) explained that data technology, which is likened to the crude oil of the twenty-first century, determines whether to secure a competitive advantage in the future and is the key to creating new business value. In addition, the Federation of Korean Information Industries (FKII) explained that data is the basis of core technologies of the Fourth Industrial Revolution such as big data, Internet of Things (IoT), cloud technology, and artificial intelligence (AI). In addition, the FKII explained that advanced analysis and future prediction capabilities through its use are directly related to the competitiveness of a company and the competitiveness of the country. In particular, as shown in Table 1, due to the development of digital technology and data innovation, the R&D paradigm is also shifting from observation and experiment-centered research to data-centered research. For example, recently, more than 8,000 people analyzed 10PB data generated by CERN's Large Hadron Collider (LHC) and discovered the Higgs boson. In addition, there is a growing movement to redefine the values, goals, and strategies of R&D through the opening up of R&D processes and outcomes, centering on major countries. For example, in November 2021 UNESCO adopted the Open Science Recommendations.

According to Beagrie and Houghton (2014), it was explained that major developed countries expect great economic benefits by minimizing duplicated research and increasing investment efficiency through the opening and sharing of research data. As well, Beagrie and Houghton (2014) explained that the opening and sharing of research data will activate interdisciplinary convergence and joint research and create new technologies and new industries in the future. However, since many researchers regard research data produced through national projects as their own, the openness and sharing of research data is very insufficient. In addition, since many researchers directly manage research data in personal storage devices, a large amount of research data is lost without being systemati-

Table 1. Research paradigm shift

1st Generation Research Environment	2nd Generation Research Environment	3rd Generation Research Environment	4th Generation Research Environment
Experience-based research	Theory-driven research	Computer resource-focused research	Data-driven research
After collecting and producing data through observation or experiment, research is carried out based on it	Research has been conducted through modeling and generalization methods that have been conducted for hundreds of years	Research on simulating complex phenomena using computing resources in a way that has been possible over the past decade	Research that uses vast amounts of data as a central tool for research
			

 https://doi.org/10.1633/JISTaP.2022.10.S.3

cally managed. To solve this problem, some developed countries are implementing policies related to research data. In 2013, the United States promulgated guidelines for the management and sharing of research data by the White House Office of Science and Technology Policy. And through this, access to research results (research publications, digital data) carried out with public research funding of more than $100 million annually through public research management institutions is being improved. France promulgated the Digital Republic Act in 2016. Therefore, researchers are obligated to open and freely share research papers or research data that received 50% or more of its research funding from public funds in France or the European Union (EU). In addition, countries implementing research data policies are building and servicing research data platforms that support guidance, education, inter-researcher cooperation, and portal services for data management and sharing at the national level. In Korea, research data is defined in the national R&D information processing standards in the National R&D Innovation Act. In addition, research data management standards are announced. In accordance with these policies, a research data platform has been built since 2018 and operated since 2020.

This paper describes the National Research Data Platform (DataON), which has being built and serviced by the Korea Institute of Science and Technology Information. DataON is a platform that systematically manages research data produced by Korean-funded research institutes at the national level. In addition, DataON supports connection services with leading overseas research data platforms such as Europe's OpenAIRE (https://www.openaire.eu/), Australia's ARDC (https://codata.org/), and Japan's IRDB (https://irdb.nii.ac.jp/en/).

2. RELATED WORKS

This chapter introduces the current status of major global research data platforms linked with DataON.

First, OpenAIRE. This is the world's largest research data portal service. It collects research data produced from projects supported by the EU and provides them to researchers. Research institutes from many EU countries operate the service by being supported by EU project funding. It provides researchers with about 2.48 million research data in all fields of science and technology.

The architecture of OpenAIRE is largely composed of a connection part for data provision, a data purification and storage management part, and a service part, as shown in Fig. 1. The link part processes metadata, original text data, research data, etc. provided by literature repositories, journal materials, and data repositories. Within the

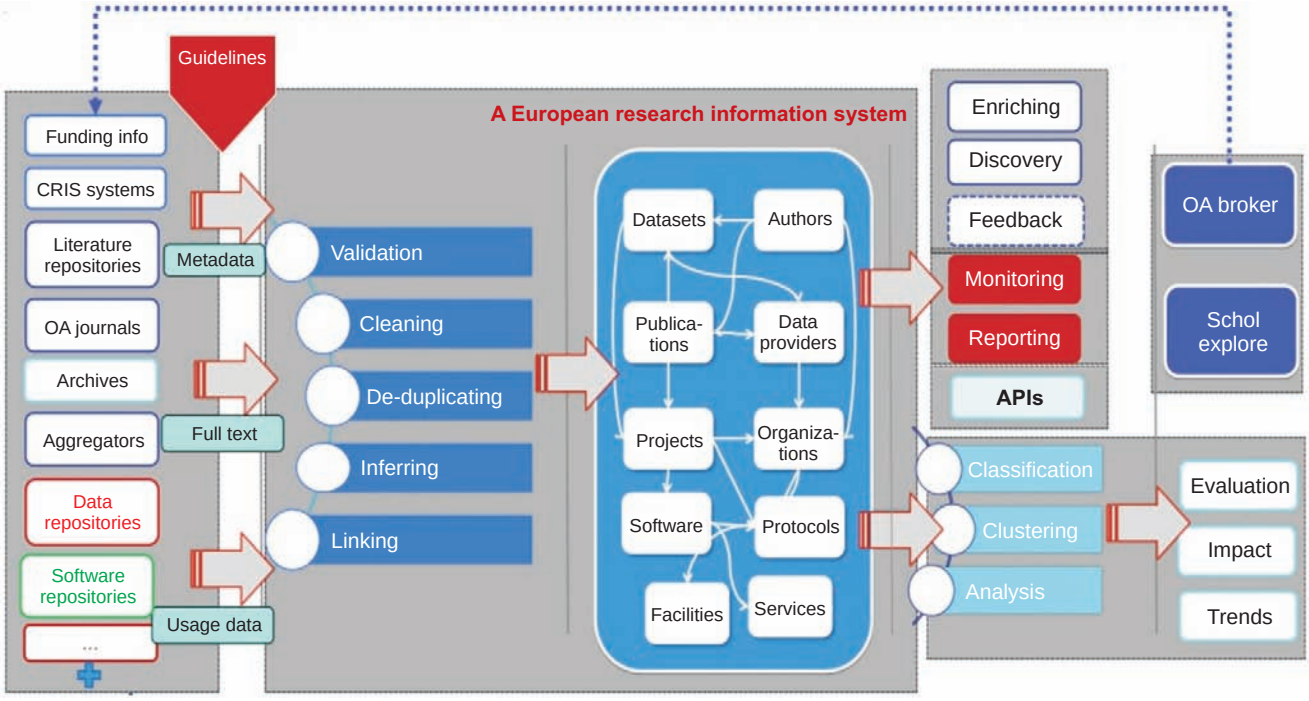


Fig. 1. Architecture of OpenAIRE (Open Access Infrastructure for Research In Europe).

platform, data validation, purification, deduplication, and link processing between metadata and similar metadata or metadata and related data are performed. Based on this, it provides search, crowdsourcing, monitoring, and reporting services, and provides services such as data evaluation and impact analysis through classification, clustering, and analysis processes.

The difference from DataON is that OpenAIRE provides only the metadata of research data, whereas DataON provides some raw data and data analysis environment.

Second, ARDC. ARDC collects Australian research data and provides it to researchers. It provides raw data and analysis environment through connection with infrastructure organizations in Australia. It is operated by the Australian Research Data Commons (ARDC). About 300,000 cases of research data in all fields of science and technology in Australia are provided to researchers. The data service of ARDC is called ANDS (Australian National Data Service).

The architecture of ANDS is largely composed of a development framework, various utilities, metadata collection and transformation management, and data utilization management, as shown in Fig. 2. The development framework provides services with a data management system that can be easily understood by government institutions, research institutions, and research investment institutions, and provides various utility services such as search, collection registration, and permanent identifier management. In addition, it is configured to collect, transform, and manage metadata owned by the institution.

Third, IRDB. IRDB (Institutional Repositories DataBase) links and collects Japanese research data and provides it to researchers. Research papers and related research data are collected by linking repositories with universities in Japan. It has about 70,000 research data in all fields of science and technology in Japan.

It consists of research data management (GakuNin RDM), a repository (WEKO3), and search (CiNii Research) as shown in Fig. 3. The research data management part manages research data and files generated in the course of project execution, including preservation and version management of files related to research data, and the repository part manages academic data (research reports, gray literature, theses, research data, etc.). CiNii Research, a search service provided by the National Institute of Informatics (NII) in Japan, is applied to the search part.

3. KOREA RESEARCH DATA PLATFORM (DataON)

3.1. Introduction to Overall Concepts of DataON

The study by Hey et al. (2009) tells that data is the foundation of the Fourth Industrial Revolution, and as the importance of data analysis and utilization increases, data utilization capabilities are emerging as the core of scientific and technological competitiveness in the Fourth Industrial Revolution. In addition, the paradigm of R&D is rapidly shifting to being data-centric, including data generation, processing, and analysis. In addition, from the OECD's report (2015), the Open Science movement, which seeks to publicly spread the results of public research (publications and data) in digital form, to enhance its socio-economic benefits, is actively underway.

The EC (European Commission)'s report (2018) tells

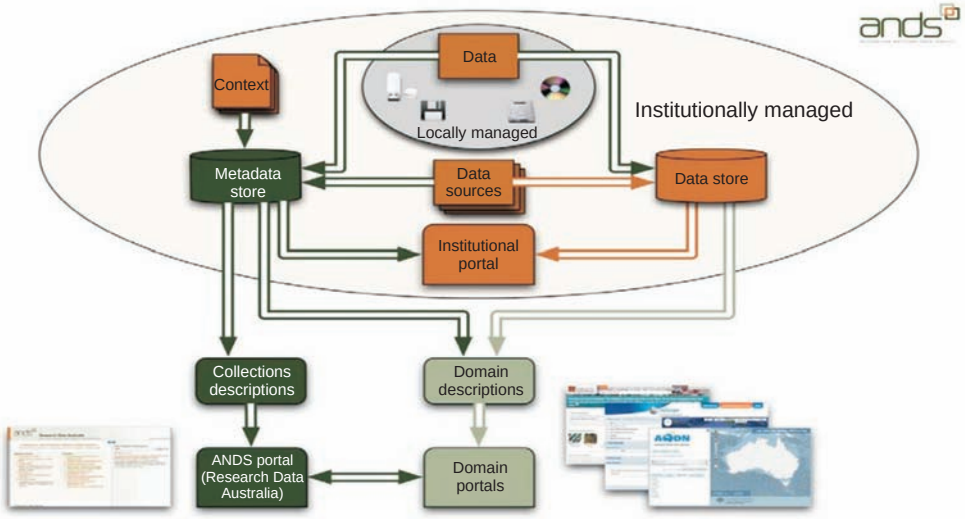


Fig. 2. Architecture of ANDS (Australian National Data Service).

 https://doi.org/10.1633/JISTaP.2022.10.S.3

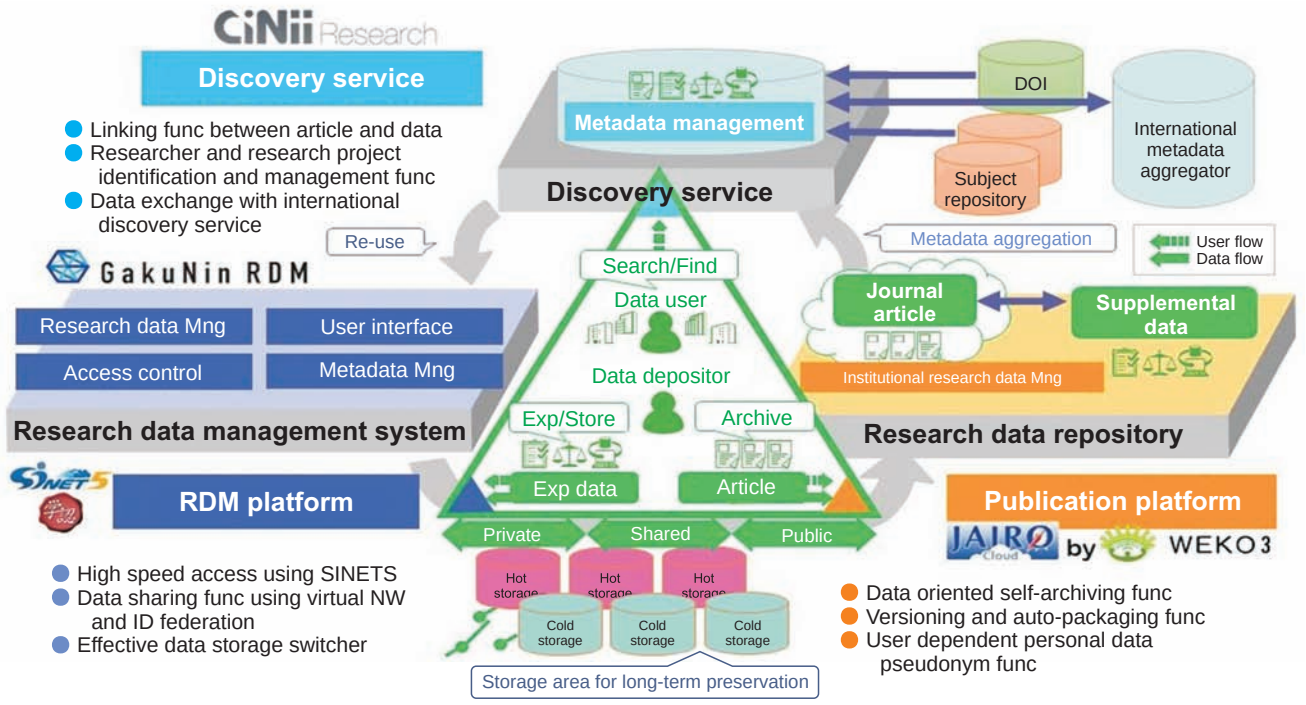


Fig. 3. Architecture of RCOS (Research Center for Open Science and Data Platform), including IRDB (Institutional Repositories DataBase).

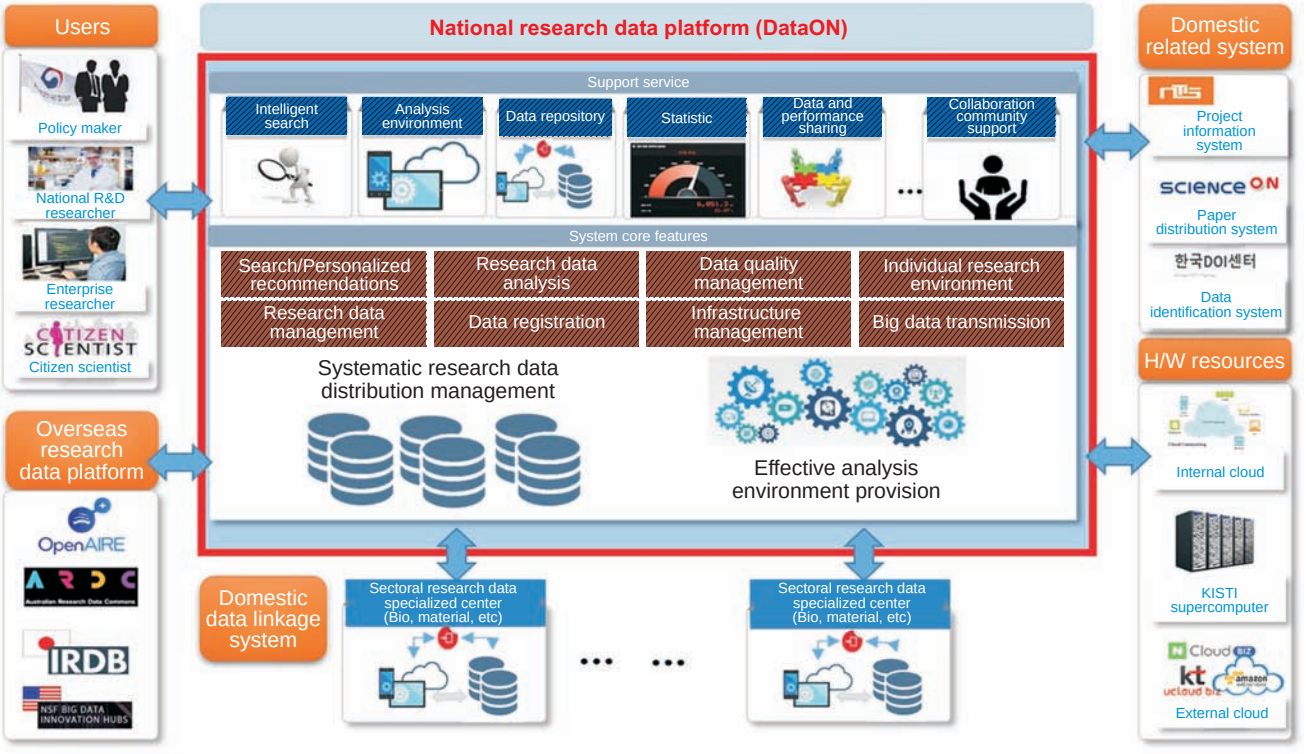


Fig. 4. Conceptual diagram of DataON.

that as the importance of data sharing increases, the FAIR (Findable, Accessible, Interoperable, and Reusable) principle for data is required as a core concept for Open Science, which enables clear access using metadata, including permanent identifiers.

While advanced countries are establishing various policies and platforms for data utilization based on open science, data management and utilization in Korea are insufficient. Accordingly, KISTI is designing and building DataON, a Korea research data platform, to systematically manage and share research data.

DataON was designed with core concepts such as open science, the FAIR principle, and research reproducibility, and the conceptual diagram of the target system is shown in Fig. 4.

DataON stores research data, which is factual data gathered through various experiments, observations, investigations, and analysis conducted in the course of conducting R&D tasks, and is essential data for the verification of research results.

To construct research data, KISTI's research data, research data from specialized centers such as biology, materials, and research data from government-funded research institutes are linked. As well, in the case of overseas data, major advanced research data platforms such as Europe's OpenAIRE, Australia's ARDC, and Japan's IRDB have been linked, and the goal is to establish additional links with overseas platforms every year. DataON systematically manages and operates the collected research data.

DataON provides useful information to policy makers, researchers of funded research institutes, corporate researchers, and citizen scientists through connection with NTIS (National Science and Technology Information Service)'s project information, ScienceON's thesis information, and the DOI system of Korea DOI Center. In addition, a data analysis environment is provided so that researchers who conduct data-based research can share/utilize collected/constructed research data. Analysis environment resources include DataON's computing resources and KISTI's supercomputers targeting huge computing resources. Also, as major services to support researchers, it provides services such as intelligent search service, analysis environment, data repository, statistics, data and performance sharing, and community cooperative research.

Based on the research lifecycle of the researcher, the functions and services of DataON corresponding to major research stages are as follows in Fig. 5.

1) In the idea derivation (Ideas) stage, the identification of research data through various searches, workflow, analysis apps, data analysis case studies from markets, and online education and content for the use of the platform are used.

2) In the stage of exploring collaborative researchers (Partners), users can utilize the community, use the collaborative researcher recommendation function, and recruit collaborative researchers through the data, app, and workflow sharing function.

3) In the proposal writing stage, data, data analysis history, and recommendations and reasoning functions for

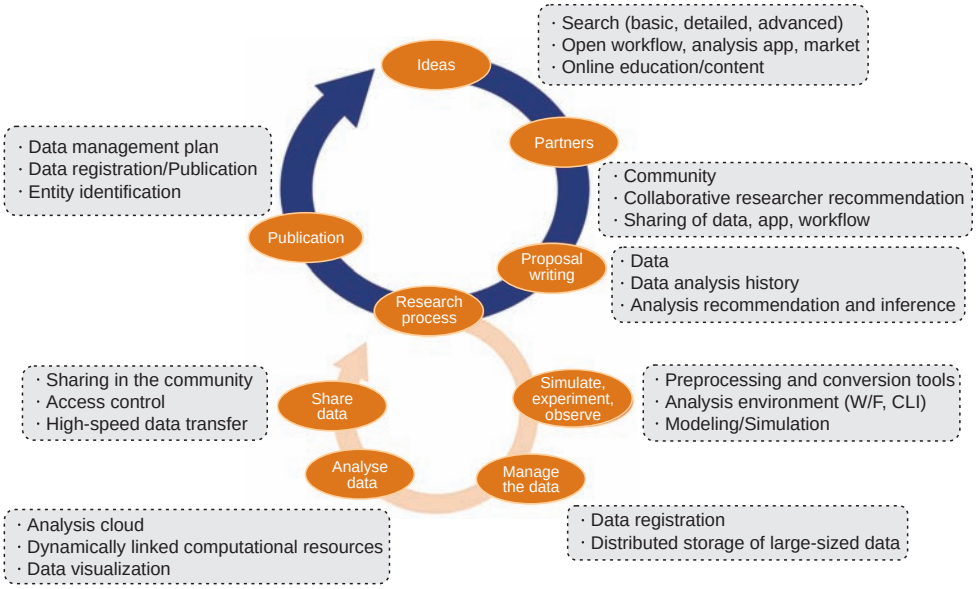


Fig. 5. Functions of DataON corresponding to the research lifecycle.

 https://doi.org/10.1633/JISTaP.2022.10.S.3

data and analysis can be utilized.

4) In the research progress stage, each detailed stage can be divided as follows.

First, in the simulation, experiment, and observation stage, functions such as preprocessing and conversion tools, workflow and CLI-based analysis environment, modeling, and simulation can be utilized. Second, in the data management stage, research data registration, data distribution, and storage functions can be utilized. Third, in the data analysis stage, functions such as dynamic computational resources and data visualization can be utilized through the connection of analysis clusters and supercomputers. Finally, in the data sharing stage, functions such as sharing with collaborative researchers and communities, access control, and high-speed data transmission can be utilized.

5) In the publishing stage, functions such as data management plan (DMP), data registration and publication, and entity identification based on Permanent Identifier (PID) can be utilized.

DataON has been designed to have the functions and technologies required throughout the research life cycle, and is being built on an annual basis.

3.2. DataON's Search

3.2.1. Search System

DataON's data search system is composed of a master node responsible for creating/changing/deleting indexes, a data node responsible for storing data, CRUD (Create/Read/Update/Delete), etc., and a coordinator node that responds to user requests.

Elasticsearch Engine is an open source real-time distributed search engine based on Apache Lucene and supports distributed search and analysis of JSON-based unstructured data. It is very convenient to install and extend the server, and it provides real-time search service support, distributed and parallel processing, and multi-tenancy functions, and various functions can be implemented and applied in the form of plug-ins. In addition, since the cluster can be configured, it is very easy to respond when the capacity of the search target increases. In addition, it is equipped with a Korean morpheme analyzer, enabling effective Korean searches.

3.2.2. Search Service

DataON's search service aims to provide various additional services to share and utilize data easily and quickly by arranging content optimally. To this end, the service was designed through UI/UX specialized consulting, and the requirements of 993 researchers in six fields, including

Table 2. Comparison of search services of domestic and foreign research data platforms

Categorization		OpenAIRE	ARDC	AI-Hub	KOBIC	DataON
Search	Integrated search	●	●	●	●	●
	Detailed search	●	●			●
	Facet search	●	●		●	●
	Map search		●			●
	Image search					●
Metadata detail view	Detailed metadata view	●	●	●	●	●
	Row file view	●	●			●
	Row file download		●	●	●	●
	Row file preview					●
	Sharing	●	●			●
	Bookmark	●	●		●	●
	Utilization statistics	●	●		●	●
Additional service	Research data graph		●			●
	Collaboration of analysis platforms					●
	Generation of citation information	●	●			●

specialized centers and research institutes, were reflected.

In addition, similar services at home and abroad were benchmarked and designed through analysis of design/convenience/layout aspects. In addition, as shown in Table 2, the purpose and nature of the research data search service, the characteristics of the provided content, and the differentiation from other platforms were set as comparative indicators to review and analyze domestic and foreign search services and functions similar to DataON.

DataON's search function is designed based on simple search, prioritizing search convenience, and additional functions such as integrated search, detailed search, map search, content search, and data provider search are configured to enable close search according to user search level, preference, and data type.

Integrated search refers to a simple keyword-based search, detailed search refers to a field-specific search function constituting a data field, and map search refers to a latitude/longitude coordinate-based search using OpenStreet Map, an open-source participatory free map. Content-specific search refers to close search by data type of dataset, table/figure, and software, which are the main contents of DataON. In addition, search by data provider refers to a function that allows data search by domestic and foreign platforms and repositories linked to DataON.

In particular, a more precise search can be performed while narrowing the search scope by applying various facets on the first screen that show the search results after executing the search. To facet refers to narrowing down the search results, and it is a search method that selects the information you need while gradually narrowing the range among numerous data. Through this, it is possible to access the desired information more quickly and precisely.

The search result detail screen can be divided into a data information area and a utilization information area, and the main information provided in each area is as follows.

1) Data Information Area
- Information of main data (main information such as title, author, description, and disclosure and access right information)
- Entity relationship information (relationship information graph between entities, such as tasks and participants, data holding organizations, and repositories)
- Linked information (project information and thesis information)
- Landing page link (data provider) or download link (raw file)

2) Usage Information Area
- Error report, recommendation/interest selection
- Data inquiry/recommendation/share aggregate information
- Source, repository, DOI information
- Citation information creation/copying (citation information generated according to the citation style of major journals such as *IEEE*, *Nature*, or *Science* through CSL [Citation Style Language])
- License
- Share (share social network service such as Facebook, Tweeter, Community, or User Share)

As described above, in the detailed screen, various information for sharing and utilization is provided along with detailed information about the data, thereby increasing the usability of data.

3.3. DataON's Data Linkage

3.3.1. Metadata Schema Design

For the metadata schema design, schemas of major affiliated organizations such as OpenAIRE and ARDC are analyzed, and domestic standards (TTAK.KO-10.0976 metadata for management and sharing of research data, Korea Information and Communication Technology Association) and foreign standards (Data Catalog Vocabulary), DataCite, and Schema.org were used as reference

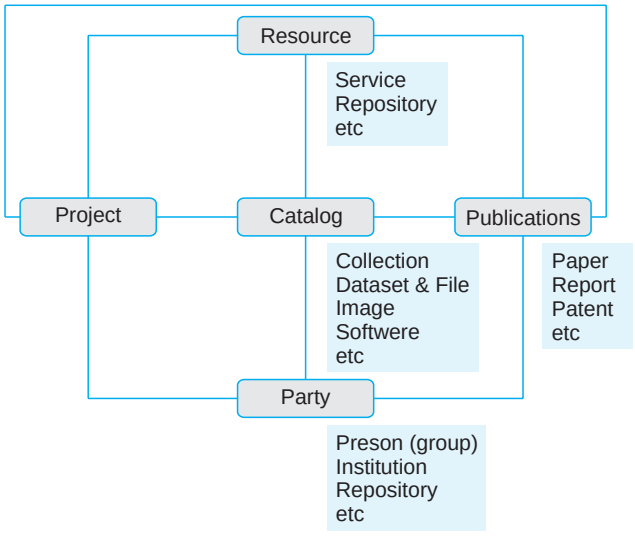


Fig. 6. Metadata schema object configuration and relationship diagram.

 https://doi.org/10.1633/JISTaP.2022.10.S.3

standards to design DataON's metadata schema.

The metadata schema is defined as a total of five entities: Project, Resource, Catalog, Party, and Publication, and the composition and relationship graph of each entity is shown in Fig. 6.

Entity attribute definitions and attribute domains (value range, data type, and restrictions) are defined, and resource types (11) of instances, identifier types (35), and relationship types (26) between instances are additionally defined.

All entities were designed based on PID. Project Entity defined 29 properties to store task information. Resource Entity defined 18 properties to store information such as repository and service. Catalog Entity defined 38 properties to store various classifications of research data (dataset, image, file, report, software, workflow, model, etc.). Party Entity defined 17 properties to store information such as individuals, groups, repositories, and institutions, while Publications Entity defined 20 properties to store information such as papers, patents, and reports.

3.3.2. Data Linkage Method

DataON and related organizations link data in the form of automatic connection by API (OpenAPI/OAI-PMH), data registration by direct data input by the person in charge of the related organization, and data registration by bulk files, such as CSV. The linkage process is carried out through 1) linkage consultation, 2) distribution of metadata schema of DataON, 3) registration of API (OpenAPI, OAI-PMH) interface, 4) mapping of schema to DataON of linkage agency, and 5) data linkage. However, if the interface of the affiliated organization is not established, it is registered through the metadata file or by directly entering the DataON.

The data manager of each organization is in charge of overall practices (registration/collection/embargo/authority management, etc.) related to the connection between the organization's repository and DataON, and has roles such as schema management (inquiry/registration/collection/modification/deletion) authority, schema mapping (inquiry/registration/modification/deletion) management, and various statistical information inquiries.

3.3.3. Metadata Construction and Linkage Status of DataON

Building and linking metadata is in progress by establishing a plan to link domestic and foreign platforms on an annual basis. Table 3 shows the status of linkage and data construction so far.

3.4. Data Analysis Environment of DataON

DataON's analysis environment provides a one-stop analysis environment based on research data, such as data, software, and images built on the platform, which helps to

Table 3. Current status of data linkage by platform/institute/repository

Categorization	Institute/Platform/Repository	Starting year of connection	Connection method	Number of data
Domestic	KISTI (Korea Institute of Science and Technology Information)	2018	Registration	313
	KIGAM (Korea Institute of Geoscience and Mineral Resources)	2020	RESTful	2,792
	KRISS (Korea Research Institute of Standards and Science)	2019	File	59
	NIA (National Information Society Agency)	2018	File	21
	GSDC (Global Science experimental Data hub Center)	2018	File	4
	KOPRI (Korea Polar Research Institute)	2021	OAI-PMH	29,303
	KCRC (Korea Carbon Capture & Sequestration R&D Center)	2020	Registration	170
	NRMS (National R&D Reports Management System)	2019	File	1,871,818
	AIDA (https://aida.kisti.re.kr)	2020	RESTful	10
	Individuals	2018	Registration	144
Overseas	OpenAIRE (Open Access Infrastructure for Research In Europe)	2018	OAI-PMH	963,088
	ARDC (Australian Research Data Commons)	2019	OAI-PMH	190,572
	IRDB (Institutional Repositories DataBase)	2020	OAI-PMH	75,231
etc.	COVID-19	2019	Registration	392

reduce the time required for R&D. In addition, it can be used as a tool to support research reproducibility that can verify research data and apps registered by other researchers. DataON's analysis environment consists of front-end web interfaces for users such as workflow and JupyterLab, back-end resources for computing, APIs connecting front-end and back-end, and diagrams of key components. These are shown in Fig. 7.

For the user environment of the analysis environment, two environments were designed: a workflow-based analysis environment for beginner data analysts or users studying data analysis, and a JupyterLab-based analysis environment for data analysis experts.

The system environment of the analysis environment consists of a Kubernetes cluster for computational resources, user data, apps, Lustre storage for storage in the development environment, Docker Registry for container image storage, multiple users, and FreeIPA for user information and authentication.

Users who have obtained access to the analysis environment can analyze data while creating a workflow on the canvas, or by coding with JupyterLab through a multi-user access framework called JupyterHub.

According to online documentation of JupyterHub, JupyterHub is a good way to provide Jupyter Notebook (Lab) to multiple users, and it can be used in various ways in student classes, data science groups, or scientific research groups. It is a multi-user hub that creates, manages, and proxies multiple instances of each Jupyter Notebook server. The workflow-based analysis environment is a drag and drop analysis environment that creates and executes ML (Machine Learning) or DL (Deep Learning) workflows using a canvas. The main procedures are as follows.

1) Drag and drop data and apps onto the canvas.
2) Connect data and apps.
3) Save and run the workflow.
4) Execution results are saved in the location set in the app.

The JupyterLab-based analysis environment refers to an analysis environment through a web-based interactive development environment. According to online documentation of JupyterHub (2022), JupyterLab is a web-based interactive development environment that supports more than 40 programming languages, including Python and R, and provides interactive output, visualization, and documentation through Markdown.

The computational resources for this user environment are allocated in Pod units that collect Linux containers in the Kubernetes cluster. Pod is also divided into CPU or GPU resources, and parallel distributed processing such as MPI (Message Passing Interface), Horovod, and IPyP-

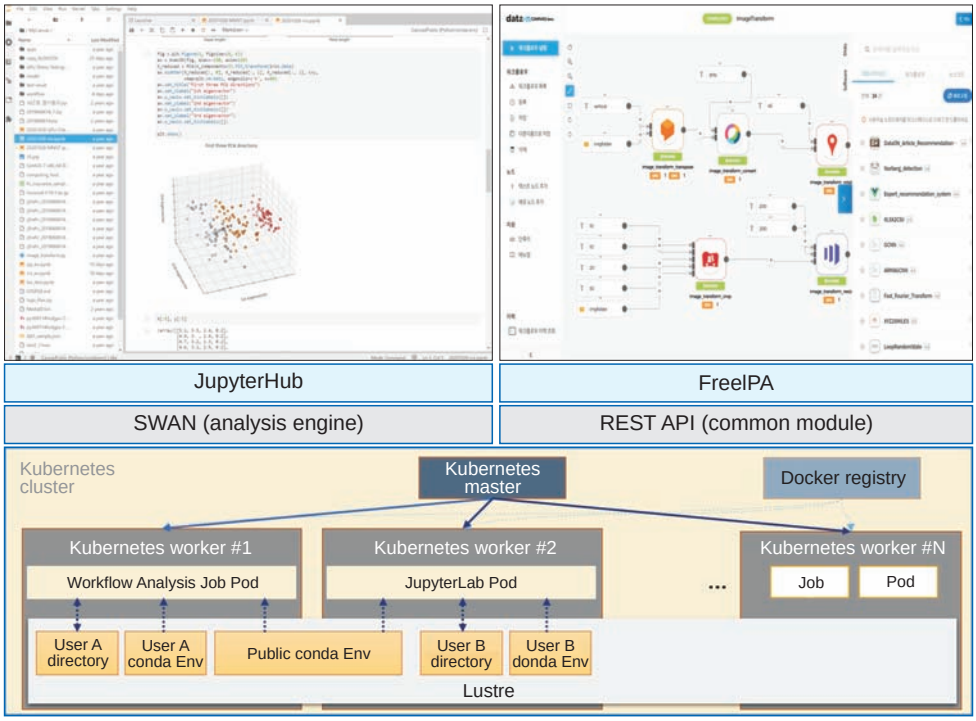


Fig. 7. Configuration diagram of DataON's analysis environment.

 https://doi.org/10.1633/JISTaP.2022.10.S.3

arallel is made possible through multiple Pods. As well, it was designed based on Conda environment to share users' execution environments and development environment in app registration, workflow, and JupyterLab environment. Through Conda, users can create an independent virtual work environment for each project and specify a separate library module, version, etc. for each project. In this way, Conda allows installing and running of multiple projects independently of each other on one's own Pod. Also, by sharing the app and the Conda virtual environment created in this way, it is possible to increase the reproducibility of research in which other users run the app.

In order to share these data, apps, workflows, and Conda virtual environment, it was implemented through Docker Image, API, etc. based on the Lustre storage system, which is a high-speed mass storage device.

3.5. Status of DataON's Infrastructure

DataON was designed with a Micro Software Architecture structure (homepage portal, data management, analysis environment) based on open source, and was built based on the e-government framework. The configuration of DataON's software and hardware infrastructure is as follows in Fig. 8.

The representative hardware equipment constituting DataON consists of a server cluster, data storage, backup storage, and network. In the case of the server, it is used to provide Web, DB, search service, and computational resources to DataON users based on x86 servers (total 48 units). In the case of analysis equipment, it consists of nine resources for AI learning and inference, and there are three NVIDIA V100 GPUs per server. Data storage is being built with the goal of 2PB based on the Lustre file system and is being used in the data storage and analysis service environment. The backup storage was built for archiving research data and for backing up the DataON system, and is being built with the goal of 1.4PB. For the network, a high-speed network (10Gbps network, 100Gbps-based infiniband network, etc.) was built to transmit large-scale research data. Most servers and data storage have improved system stability and availability through redundancy or cluster construction. In this way, software was built based on high-performance and stable hardware.

DataON is built based on open software, and the main software is web servers (Apache HTTPD), WAS (Web Application Server; Apache Tomcat), database servers (PostgreSQL), search servers (Elastic Search Engine), OpenStack for Hadoop/Spark, IDR hosting, Kubernetes for computational clusters in the analysis environment, and FreeIPA for authentication and access control. Further, for service availability, the web server is physically duplicated and the load is distributed through a load balancer. The WAS server is duplicated using mod-jk, an Apache module that connects the web server and the Tomcat servlet container. PostgreSQL database cluster has improved availability and performance through Pg-Pool. As such, it was designed and implemented with the goal of improving service continuity and operability through physical hardware multiplexing and software design and implementation.

4. INSTITUTIONAL DATA REPOSITORY (NaRDA)

The institutional data repository (IDR) is a system that provides functions for registering, storing, posting (publishing), and searching research data generated in scholarly activity. Research data generally defines data produced in the course of experiments, investigations, and analysis to prove research. The IDR is provided as a web service so that users can search for data through the Internet. We developed IDR to make a national research data governance and to spread IDR to Korean research institutions toward the encouragement of establishing an open science policy. We call the IDR the National Research Data Archive (NaRDA in short). NaRDA provides the following main

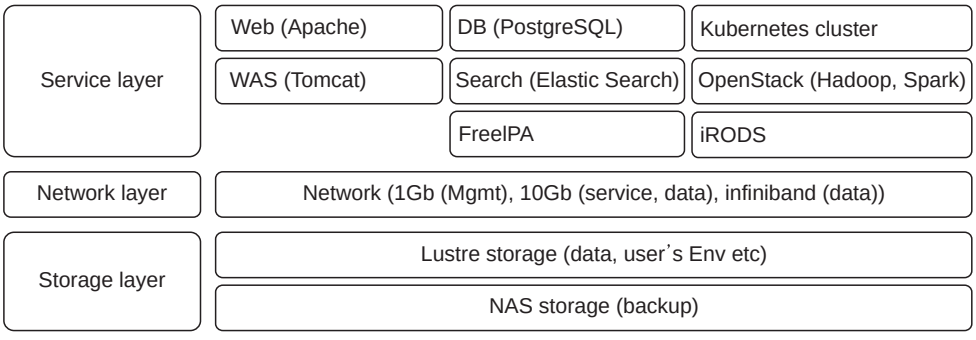


Fig. 8. Configuration diagram of DataON's software and hardware.

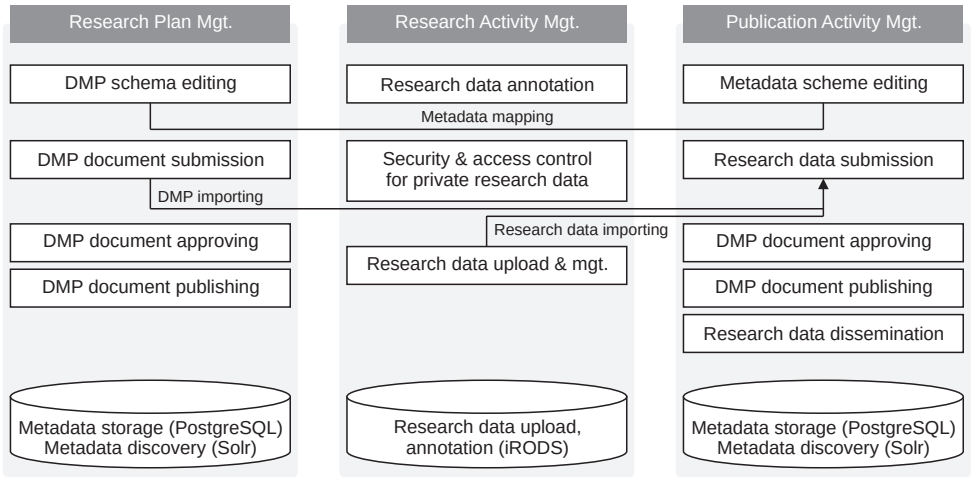
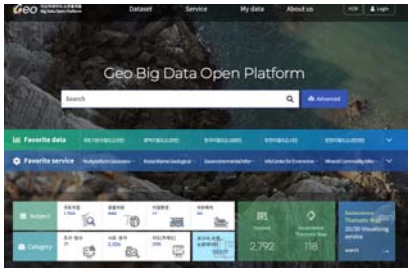
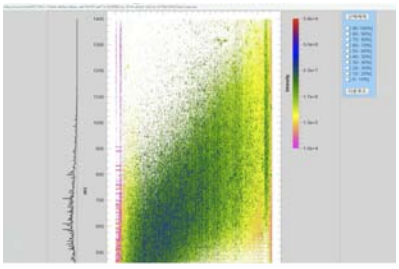


Fig. 9. Conceptual architecture for supporting overall scholarly activities (DMP to research data publication). DMP, data management plans.

Table 4. Examples of customizing NaRDA

Geo Big Data Open Platform (KIGAM)	KIOM Oriental Medicine Repository	Visualizing Mass Spectrum (KBSI)
		

NaRDA, National Research Data Archive; KIGAM, Korea Institute of Geoscience and Mineral Resources; KIOM, Korea Institute of Oriental Medicine; KBSI, Korea Basic Science Support Institute.

features as in Fig. 9. First, it provides a function to identify and authenticate users who register, search, and download research data. For this, the system authenticates and identifies users who register and manage data through federated user authentication such as ORCID (https://orcid.org/) and KAFE (https://www.kafe.or.kr/). Second, it stores and manages all data generated from the researcher's overall scholarly activities in one integrated system. As a result, NaRDA supports all stages from research DMP to registering research data. A researcher establishes and prepares a DMP to produce data in the research planning process.

To support these activities, NaRDA provides functions for editing DMP schemes, writing DMP, submission involving review and approval processes, and publishing DMP documents. DMP has the role of informing other researchers and data storage managers in advance of what kind of data to produce and in what amount. DMP document can be opened for public or for internal use. After DMP is opened, researchers can freely upload and utilize data for research activities. For supporting users, NaRDA provides a space to upload data for each researcher account and provides a function to share data with colleagues who are conducting research together. When the research results are produced and data that can ultimately prove the research results is produced, NaRDA provides the functions to register, review and approve, and publish research data. Research data can include not only the final results, but also measurement data and analysis data used to derive the results.

Third, research data can be continuously updated and linked to other Internet resources such as publications and software. To this end, NaRDA provides a function to update data and to record the data provenance. According to the relation definition provided by DataCite (DataCite Metadata Working Group, 2021), NaRDA also provides a function to establish a relation with other Internet resources. Finally, the registered research data provides a function to link with other systems (e.g., DataON) in the form of OAI-PMH or RESTful API as well as web pages. NaRDA, in addition to providing these functions, is cur-

rently being distributed and used by 21 domestic public institutions. Representative use cases of NaRDA are as follows. The Korea Institute of Geoscience and Mineral Resources (KIGAM) has customized and utilized NaRDA to store, manage, and publish geological data. The Korea Institute of Oriental Medicine (KIOM) stores and manages oriental medicine data as a repository. The Korea Basic Science Support Institute (KBSI) stores proteomic mass spectrometry data and uses it as an additionally expanded data visualization tool for data quality management. Table 4 shows these cases of customizing and utilizing NaRDA.

5. APPLICATIONS OF DATAON

As of the end of last year, DataON is providing research data of 33,000 datasets from nine institutions and overseas data of about 1.13 million metadata sets in Europe, Australia, and Japan. Various use cases such as cloud movement prediction, road image object recognition, fruit sugar content prediction, cerebral cortex characteristic analysis, and port air pollution analysis are being presented by using data stored in DataON such as AI, 3D, IoT, brain images, and drones. Particularly, 400GB of human body image data has been used to implement a virtual body with 3D technology through technology transfer to Anatomage, a Silicon Valley company in the US this year (Fig. 10). It will be used as an educational surgical and anatomical simulation tool for researchers.

6. CONCLUSIONS AND FUTURE STUDY

DataON is a platform that collects and integrates research data generated from national R&D and provides researchers with an integrated data search and analysis environment. DataON not only collects and integrates data generated from domestic research institutions and research projects, but also provides opportunities for researchers to use research data from around the world for integrated search and research by linking with European OpenAIRE, Australian ARDC, and Japanese IRDB research data platforms. The aim of DataON is that it become the national research data governing system which manages the overall research data of South Korea. DataON needs to be connected to other domestic institutional data repositories. However, many institutions have not yet prepared a repository system and research data policy. Therefore, we developed and deliver the IDR system NaRDA. The research data repository is continuously being developed in compliance with the standard functions described in the Confederation of Open Access Repositories (COAR) Next Generation Repository (NGR) recommendation (Rodrigues et al., 2017). DataON is currently preparing for an international research data qualifying certification called CoreTrustSeal to ensure reliability as a research data platform.

In the future, we plan to conduct research and development in the following directions. With the adoption of UNESCO Open Science Recommendations (UNESCO, 2021) in 2021, the importance of open sharing of research data as well as analytical tools and infrastructure is growing. Accordingly, leading overseas research data platforms are preparing for this. In addition, we plan to develop an integrated system that can manage analysis resources and applications used inside each research institute, and establish a KRDC (Korea Research Data Commons) infrastructure that can connect and share systems such as DataON with researchers who want to use it.

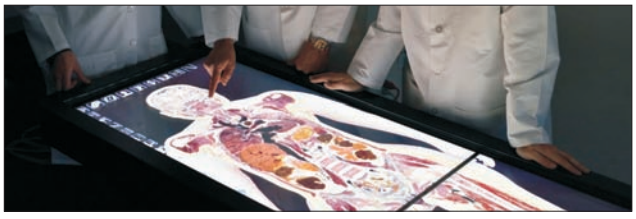


Fig. 10. Anatomage's virtual anatomical table.

ACKNOWLEDGMENTS

This work is supported by project No. K-22-L01-C03-S01, funded by the Korea Institute of Science and Technology Information.

CONFLICTS OF INTEREST

No potential conflict of interest relevant to this article was reported.

REFERENCES

Beagrie, N., & Houghton, J. (2014). *JISC report. The value and impact of data sharing and curation: A synthesis of three recent studies of UK research data centres.* Economic and Social Research Council.

DataCite Metadata Working Group. (2021). *DataCite Metadata Schema documentation for the publication and citation of research data and other research outputs. Version 4.4. Data-*

Cite e.V. DataCite Schema.

European Commission. (2018). *Final report and action plan from the European Commission Expert Group on FAIR data*. *Turning fair into reality.* European Commission.

Hey, T., Tansley, S., & Tolle, K. (2009). *The fourth paradigm: Data-intensive scientific discovery*. Microsoft Research.

JupyterHub. (2022). *What is the Jupyter notebook?* https://jupyter-notebook-beginner-guide.readthedocs.io/en/latest/what_is_jupyter.html.

OECD. (2015). *OECD science, technology and industry policy papers no. 25. Making open science a reality.* OECD Publishing.

Pilat, D., & Fukasaku, Y. (2007). OECD principles and guidelines for access to research data from public funding. *Data Science Journal*, 6, OD4-OD11. http://doi.org/10.2481/dsj.6.OD4.

Rodrigues, E., Bollini, A., Cabezas, A., Castelli, D., Carr, L., Chan, L., Humphrey, C., Johnson, R., Knoth, P., Manghi, P., Matizirofa, L., Perakakis, P., Schirrwagen, J., Selematsela, D., Shearer, K., Walk, P., Wilcox, D., & Yamaji, K. (2017). *Next generation repositories: Behaviours and technical recommendations of the COAR Next Generation Repositories Working Group*. Confederation of Open Access Repositories.

UNESCO. (2021). *UNESCO recommendation on open science*. https://unesdoc.unesco.org/ark:/48223/pf0000379949.