Choi, M-S and Lee, S. 2020. Research Data Management Status of Science and Technology Research Institutes in Korea. *Data Science Journal,* 19: 29, pp. 1–11. DOI: https://doi.org/10.5334/dsj-2020-029

REVIEW

Research Data Management Status of Science and Technology Research Institutes in Korea

Myung-seok Choi and Sanghwan Lee

Research Data Sharing Center, Korea Institute of Science and Technology Information, KR

Corresponding author: Sanghwan Lee (sanglee@kisti.re.kr)

Recent advances in digital technology and the data-driven science paradigm has led to a proliferation of research data, which are becoming more important in scholarly communications. The sharing and reuse of research data can play a key role in enhancing the reusability and reproducibility of research, and data from publicly funded projects are assumed to be public goods. This is seen as a movement of open science and, more specifically, open research data. Many countries, such as the USA, UK, and Australia, are pushing ahead with implementing policies and infrastructure for open research data. In this paper, we present survey results pertaining to the creation, management, and utilization of data for researchers from government-funded research institutes of science and technology in Korea. We then introduce recent regulations stipulating a mandated data management plan for national R&D projects and on-going efforts to realize open research data in Korea.

Keywords: open science; open research data; data management plan; data repository; research data platform

Introduction

Due to recent advancements in digital technology, such as high-tech research equipment, sensors, and data-processing technology, a large volume of research data is currently being created, and the science paradigm is also becoming more data-driven (Hey 2009). Moreover, open science, defined as efforts to make publicly funded research results publicly accessible, is now high on the global R&D policy agenda (OECD 2015). The OECD is actively promoting open science, and many countries are implementing policies and infrastructure projects related to open science (OECD 2015; RS 2012; Salmi 2015). Open science movements are expanding from open access publications to open research data. The goal of open research data is to facilitate easy access and reuse of data collected and created throughout the research process (OECD 2015). It serves as a catalyst for cooperative data-driven research to solve social problems. Open research data policies, including mandatory data management plans, have been implemented globally (Holdren 2013; ORDF 2016).

Openness is a major premise of scientific research, allowing a self-correction mechanism to work, enabling new scientific discoveries, and helping to eliminate bad science by increasing the transparency of research in general (RS 2012). The benefits of EMBL-EBI (European Bioinformatics Institute) data and services are estimated to be worth £1 billion per annum worldwide, also having a return on investment in R&D estimated to be worth some £920 million annually (Beagrie & Houghton 2016). In addition, open research data can deliver benefits to diverse members of society, such as citizens and companies, as well as researchers.

Recently, awareness of a reproducibility crisis of scientific research has increased (Baker 2016). Only six studies out of 53 'landmark' papers in preclinical cancer research were reproduced (Begley & Ellis 2012). Moreover, the cumulative prevalence of irreproducible preclinical research is reported to exceed 50% (Freedman, Cockburn & Simcoe 2015). Increasing numbers of errors are found in supplementary data submitted to well-known research journals (Economist 2016). Consequently, the numbers of journals and institutions dedicated to verifying the reproducibility of the published research have been increasing. Openness of the data created and utilized in research is indispensable to verify the reproducibility of research.

In Korea, national R&D information is integrated and managed by the NTIS (https://www.ntis.go.kr), but only a limited amount of research data is being managed and shared on a nation-wide level. Infrastructure for the sharing and reuse of research data is also insufficient. In this paper, we investigate survey results pertaining to the creation, management, and utilization of research data for science and technology researchers from government-funded research institutes in Korea conducted in 2018. Subsequently, we introduce ongoing efforts to realize open research data in Korea.

Related work

Tenopir et al. (2015) conducted an online survey of data sharing and data reuse practices and perceptions for 1,015 research scientists worldwide during 2013~14, and compared these findings to baseline study from 2009/2010. From their work, they found that "not only is data sharing behavior increasing, but that researchers are also viewing the practice and the overall movement more favorably." They concluded that "scientists need access to data, and a lack of data sharing can be a major impediment to progress in science."

Barsky (2015) surveyed 100 faculty members from Engineering, Natural and Physical Sciences at the University of British Columbia in 2015 with regard to their actual practices of research data management and to discern areas where the researchers would like help. Local and external hard drives are the most popular storage types for research data, and only 19.6% of the respondents reported using external data repositories. Most researchers (66.3%) share their data by personal request only, but 79.4% feel that "sharing their data enhances reproducible and collaborative science." The areas for additional help and support were found to be storage, data management plans, metadata and DOIs.

Shearer & Furtano (2017) surveyed 43 COAR members regarding requirements in the area of research data management in 2016. In their study, 53.4% of the respondents reported that they were already collecting data, while about a third stated that they had plans to do so in the near future. One half of the respondents use the same platform to manage publications and research data, and the most common platform was DSpace, followed by Dataverse. According to the respondents, the major challenges when collecting research data in their institution were to engage researchers, manage the lack of institutional policies for RDM, and to deal with infrastructure issues related to storage and preservation.

Kim & Yoon (2017) surveyed 1,237 scientists from the Community of Science's Scholar Database for the factors that influence the data reuse behaviors of scientists. According to their findings, three main areas that must be improved are educating scientists, providing internal support, and providing external resources and support, such as data repositories.

Methods

To grasp the current status and requirements associated with research data management in Korea, a survey was carried out by the authors targeting researchers from 23 government-funded research institutes[1] under the National Research Council of Science and Technology (NST) in 2018. The survey covers the current status of the creation, management, and utilization of research data, the will to participate, and the researchers' needs (see Appendix). From August 13 to September 7 of that year, the survey was distributed via our online survey system.

Results

For this survey, 301 responses from 23 research institutes were gathered. The respondents were widely distributed in nearly all fields of science and technology, as shown in **Figure 1**. The research fields in the survey were classified based on the National S&T Standard Classification System of the Korean government.

Data creation

Various types of research data are being produced from three quarters of the R&D projects, as shown in **Figure 2**. Nearly half of the data is structured, such as numerical data. Unstructured data, such as images, videos, and texts, also account for a significant portion. In addition, 68% of the data exceeds 1 GB in size, and large volumes of research data are generated, especially in the fields of earth science, energy/resources, physics, information/communication, nuclear power, and machinery.

[1] In Korea, major government-funded research institutes belong to NST, and only two out of 25 institutions could not participate in this survey due to certain security reasons. Therefore, the participating institutions can be considered to have covered all areas of science and technology in Korea.

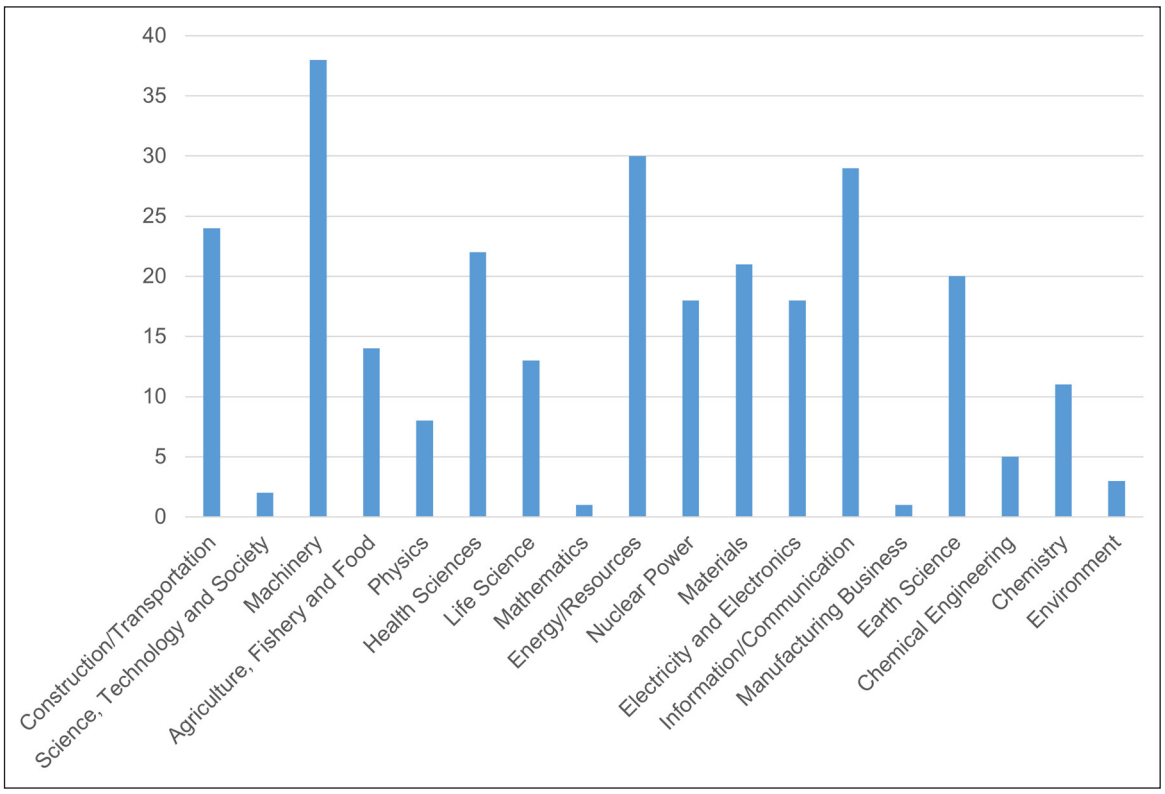

Figure 1: Research fields of the respondents.

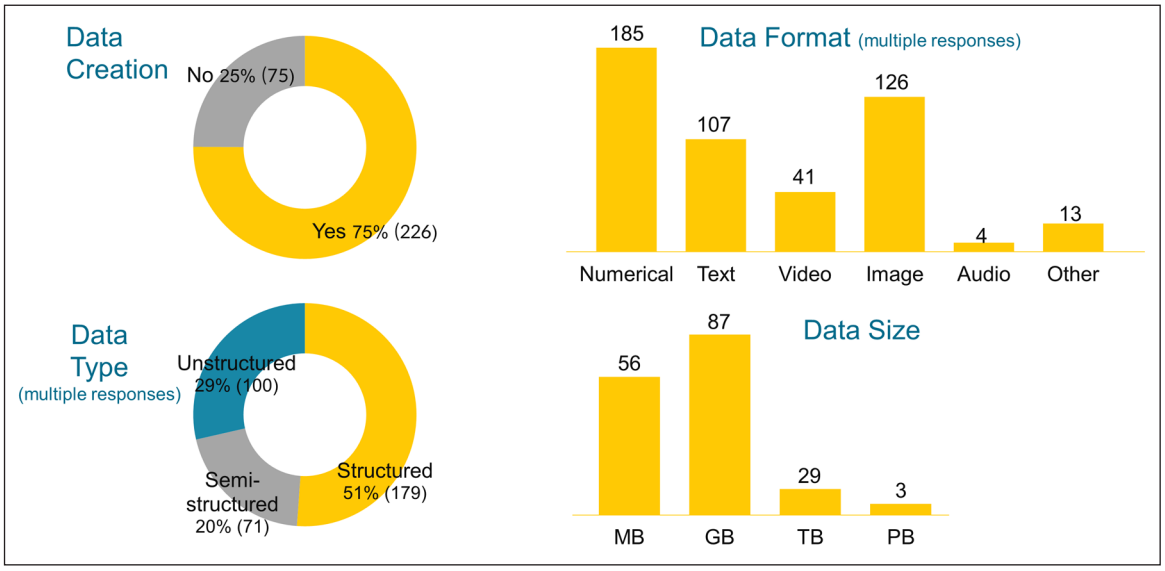

Figure 2: Status of data creation.

Data management

Although nearly all of the researchers reported that they manage their data and have data management policies or guidelines, as shown in **Figure 3**, most researchers tended to store and manage data on a personal or laboratory level, usually with their PC or via an external hard drive. According to an extra institutional interview (**Figure 4**), only two of 23 institutes manage research data at the institutional level, and three institutes have their own institutional data policy and dedicated data management department. In addition, researchers consider data management as an additional task and are sensitive to the responsibility issues of quality and reliability.

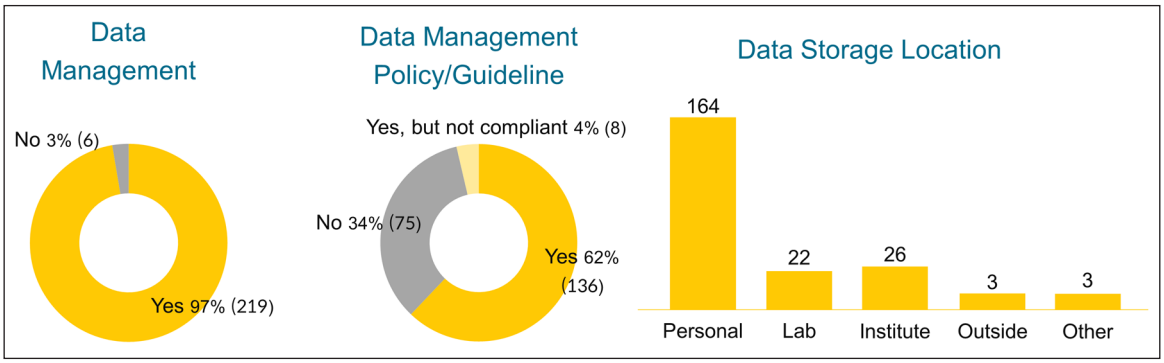

Figure 3: Status of data management.

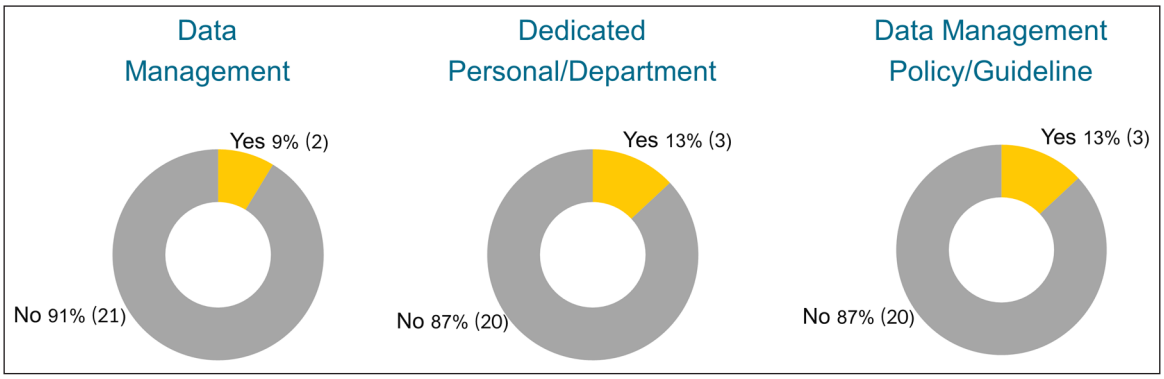

Figure 4: Status of institutional data management.

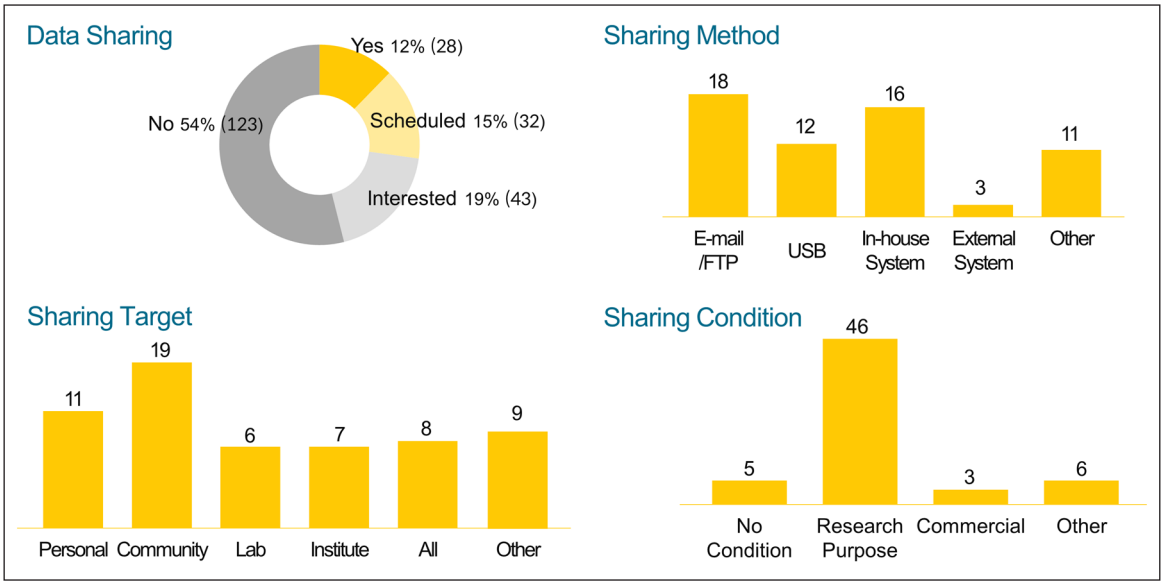

Figure 5: Status of data sharing.

Data sharing

Only a small percentage of research data is being openly shared; usually it is shared only for personal requests for research purposes, as shown in **Figure 5**. The most frequent sharing methods are E-mail/FTP and legacy in-house systems, possibly due to insufficient data-sharing infrastructure and few domestic data repositories available.

The researchers also reported that they could not share their data due to regulations and security concerns and because the data were personal and institutional assets. Insufficient incentives and insufficient infrastructure were also cited, as shown in **Figure 6**.

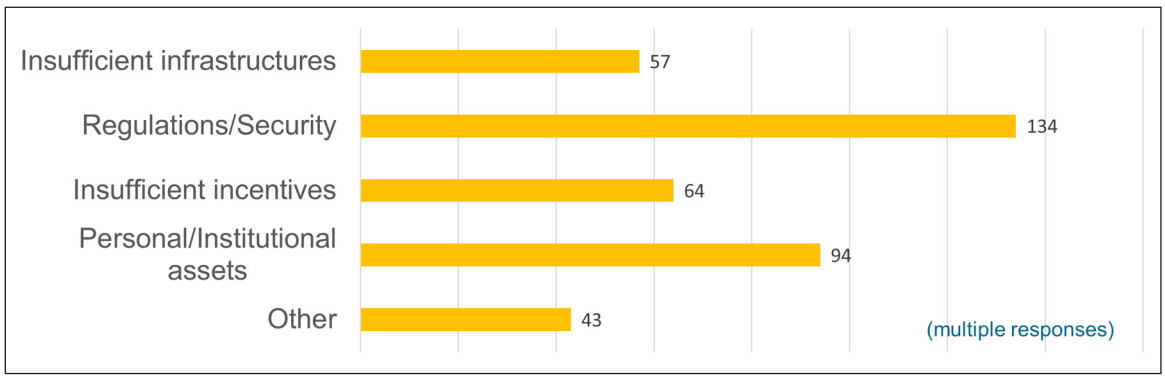

Figure 6: Difficulties in sharing research data.

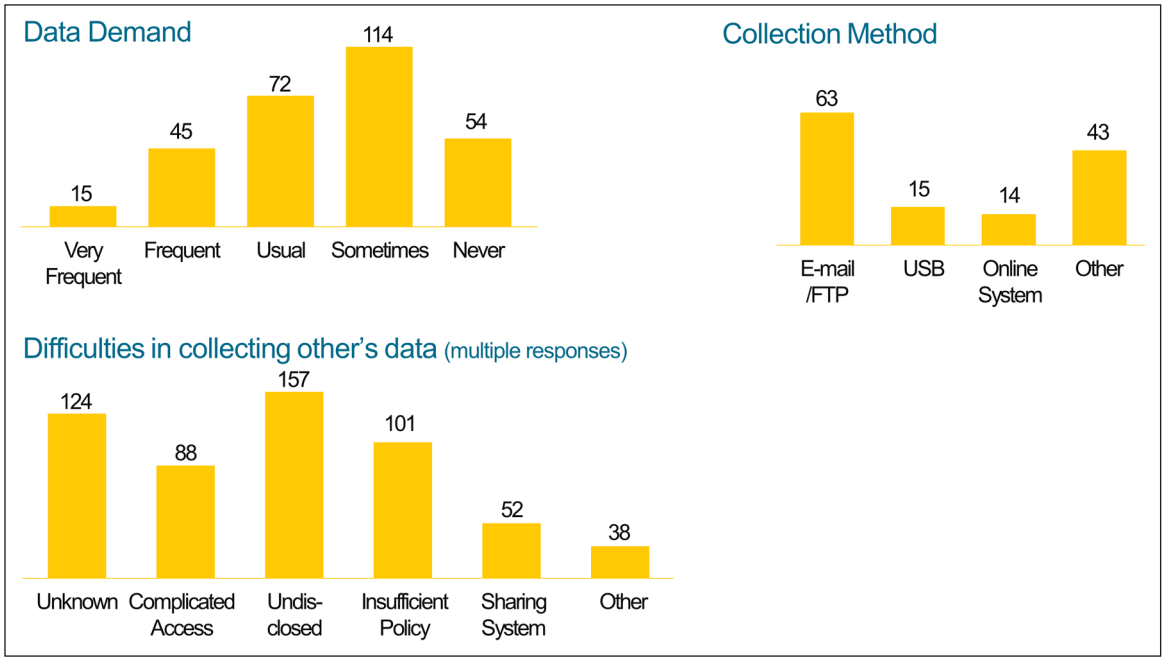

Figure 7: Status of data demand.

It was also found that 80% or more researchers have a strong demand for other researchers' data, and they usually receive such data via E-mail/FTP. The greatest hurdles when collecting data are dealing with undisclosed data, with unknown data following, as shown in **Figure 7**.

Survey summaries

Research data of various types and sizes are being created in national R&D projects. They are usually stored and managed separately on a personal or laboratory level, and are shared only through personal requests. Institutional support for the sharing and reuse of research data is still lacking, and researchers' awareness of data management and sharing is also low. As shown in **Figure 8**, 75% or more researchers create and collect research data in various R&D projects, and 80% or more researchers need other researchers' data, but only 12% at most of researchers share their data. This occurs mainly because most researchers feel that they own the data created from publicly funded projects, and the data has not yet been accepted as a major research achievement.[2]

[2] In particular, data sharing rates are considerably lower than those in other studies (Tenopir 2015, Barsky 2015). Moreover, 48% of the respondents reported that they will share their data three or more years after the end of the project. Researchers in Korea lack experience in data sharing, tend to hesitate to disclose their data, and reported that they do not receive sufficient data management support from their institutions. This tendency appears to supplement Tenopir, who stated "Respondents from Asian countries felt more strongly about data access as an important part of their own scientific pursuits." (Tenopir 2015: 19)

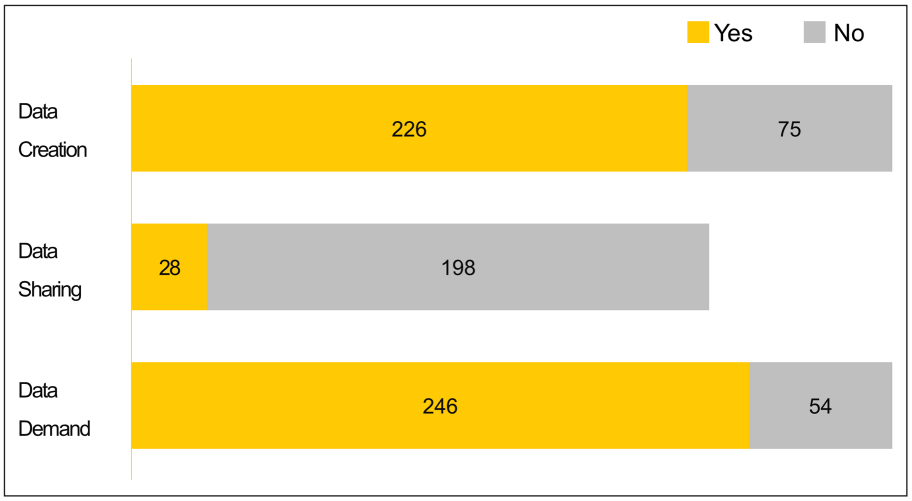

Figure 8: Gap between data sharing and demand.

In addition, researchers demanded the following to encourage data management and sharing:

- Incentivizing research data management and sharing activities so that researchers can receive proper academic credit for their contributions
- Technologically supporting data management and data-driven research by dedicated professionals, and by building institutional/sectoral/national research data infrastructure to improve the findability and accessibility of research data.
- Procuring research data quality through the entire research process
- Raising awareness of research data management and sharing for a cultural change towards open science

On-going efforts for open research data in Korea

In Korea, nearly all R&D information, including project information, participants, and outcomes, is being collectively managed by NTIS. However, only a small percentage of research data, such as biological resources and chemical compounds, is managed on a nationwide level. Data infrastructure available to individual researchers remains insufficient. To increase the reusability and transparency of research, it is a top priority to build a nationwide ecosystem of research data sharing and utilization. In Korea, there are on-going efforts to implement FAIR data principles (Wilkinson 2016) through the entire research life cycle.

National strategic plan for open research data

To build a national plan, the Ministry of Science and ICT (MSIT) organized a task force called "Collect and Renew" in 2017. They held 26 meetings over a course of five months and conducted a public hearing in December of 2017. The plan was placed on the agenda for the NSTC steering committee in January of 2018.

The plan (**Figure 9**) aims to enable national research data sharing and utilization by revising the legal system, establishing a hierarchical research data management system and infrastructure, and fostering data science experts. It consists of five key tasks: fostering a research data management system and disciplinary data sharing communities; developing a national research data platform; supporting the development of data and computing utilization experts; establishing a legal basis for research data management, sharing and utilization; and lastly promoting the industrial utilization of research data and the creation of new jobs. Currently, action plans for pursuing this strategy are being developed and several pilot projects are underway.

Mandatory data management plan

R&D projects in South Korea are conducted based on laws, specifically based on the "Framework Act on Science and Technology", the "Act on Performance Evaluation and Management of National Research and Development Projects, etc.", and "Regulations on the Management, etc. of National Research and Development Projects" (the Regulations). Based on these laws, individual government ministries operate R&D projects by establishing their own project management regulations. The government designates nine key research outcomes and systematically manages them by establishing institutions devoted to their management. These research outcomes are research papers, patents, project reports, technology summary information,

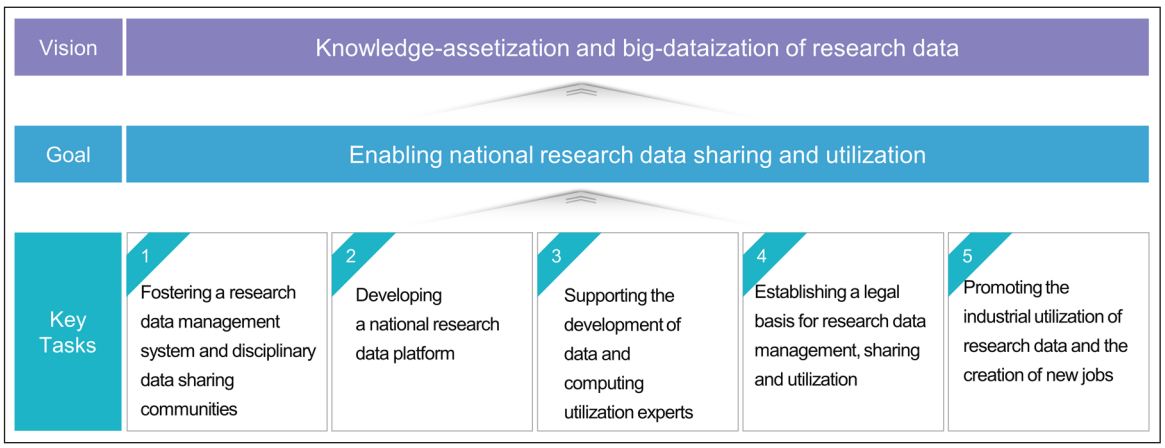

Figure 9: National strategic plan for research data sharing and utilization in Korea.

software, research facilities and equipment, chemical compounds, biological resources, and new varieties of plants.

To establish a legal basis for the collection, management, sharing and utilization of research data, which are intermediate and final products of national R&D projects, the part entitled the Regulations[3] was amended in March of 2019, going into effect in September of 2019. In this regulation, a definition of research data was added and a data management plan (DMP) became mandatory for national R&D projects.

- Definition of research data: factual records produced from various experiments, observations, investigations and analyses, etc. of research & development projects, and that are necessary to validate research findings
- Mandating DMP
  - Applying to the project in which the president of the central administrative organization admits the need for a DMP
  - Requiring a DMP for research proposals, which will be reviewed in the proposal assessment, and a request for the submittal of an updated DMP and related results with the final report

Following an amendment of the Regulations, the R&D project management regulations of individual governmental ministries are under revision. In addition, internal R&D project management regulations and guidelines were amended in R&D funding institutions and NST under MSIT. Mandatory DMP stipulations were applied to certain research projects from the National Research Foundation of Korea (NRF) and the Institute for Information & Communication Technology Planning & Evaluation (IITP) in 2019. Government-funded research institutes under NST have also been revising their own R&D project management regulations, and most of them will begin to mandate a DMP later in 2020. In the long term, a separate legislation for open research data on a national level would be more desirable.

National collaborative network

Close cooperation between various stakeholders, including data producers, data managers, researchers, policymakers, and research funders, is essential for a research data sharing and utilization system to work at the national level. Because stakeholders have little experience in managing and sharing research data and have high psychological barriers, a gradual approach is preferred through sufficient discussions along with legal and technological support. From decision-makers to research institutes, the roles and responsibilities pertaining to the establishment and promotion of research data sharing and utilization policies are defined, as shown in **Figure 10**. The establishment of a collaborative network between various stakeholders has been carefully approached. Research funders develop and enforce data policies, and the national research data center is in charge of supporting data policies, delivering data discovery services, and activating a national collaborative network. Data centers ensure proper disciplinary data management and utilization while also supporting researchers. Research institutes support intramural data sharing and utilization.

[3] There exist several acts, such as the "Official Information Disclosure Act" and the "Act on Promotion and Use of Public Data" related to public data disclosure in Korea. Because most R&D projects are controlled by the Regulations as described previously, DMP mandates are included in this regulation, which can be more effective in the situation of Korea.

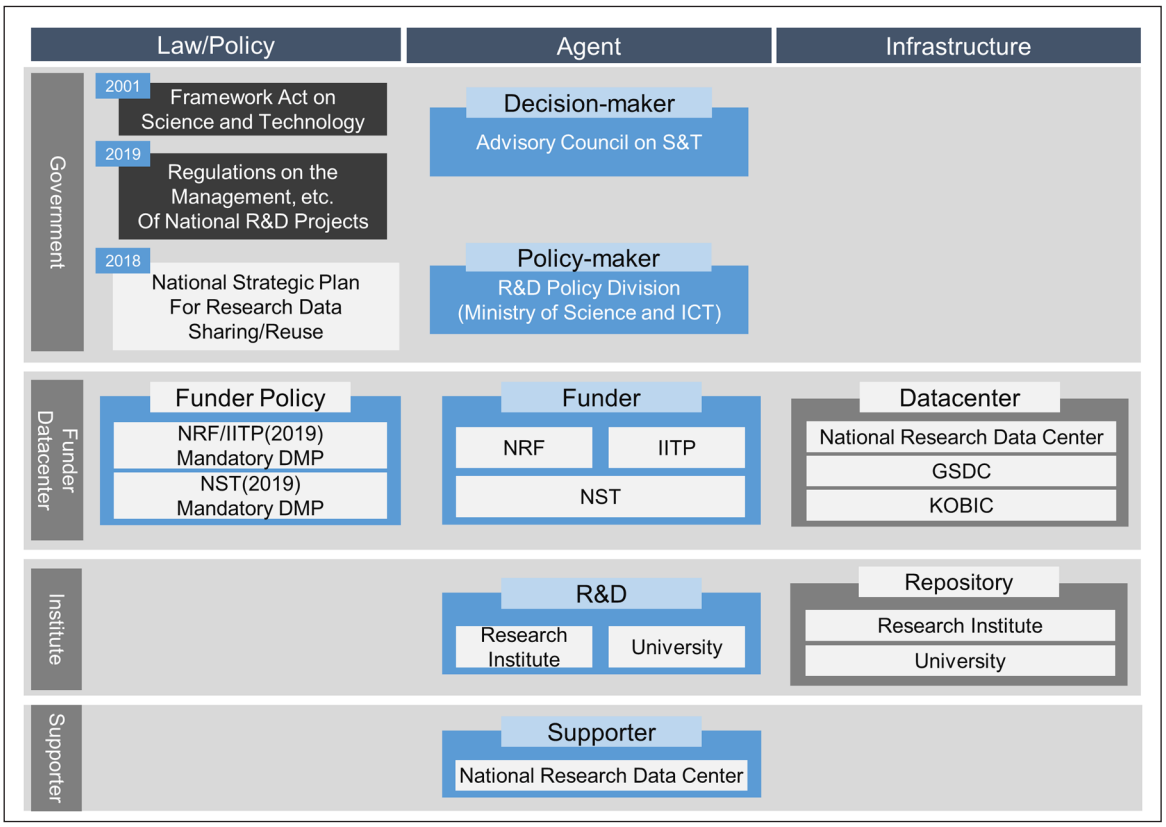

Figure 10: Stakeholders of open research data.

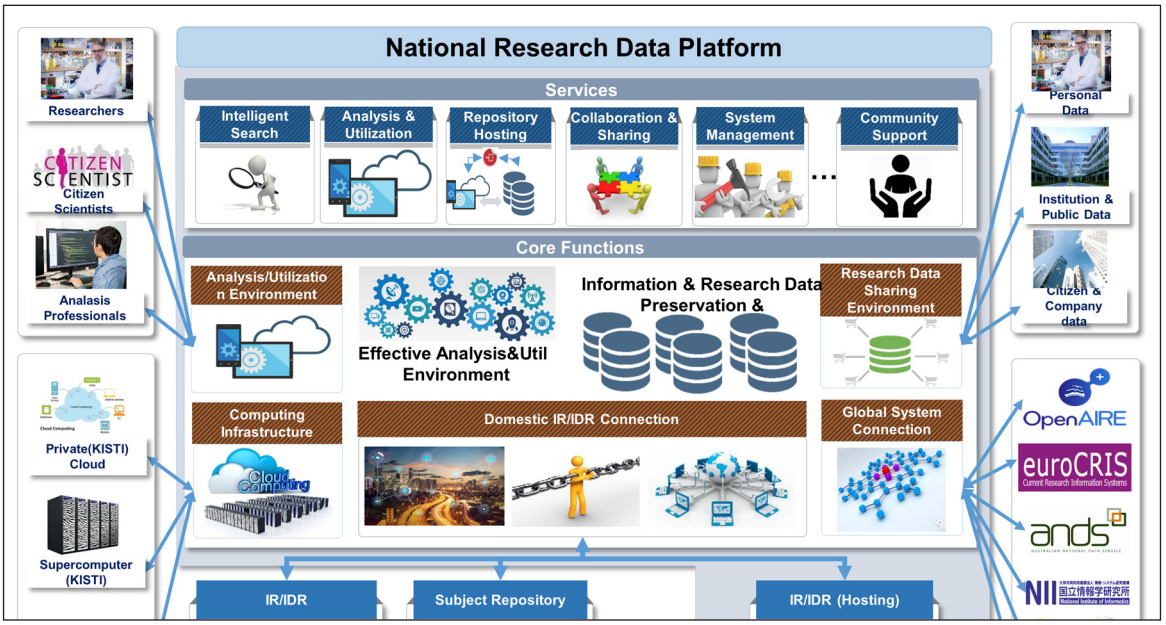

Figure 11: National research data platform.

National research data platform

A national research data platform has been in development since 2018. It will be a national research data hub with discovery, analysis and collaboration services, as shown in **Figure 11**. This platform provides functions to register and manage research data, services to connect and query domestic and international research data, cloud-based research data workflow analysis environments, research data repository hosting services, and research community services to support collaborative studies. First, the platform offers functions to store and manage various types of big research data. It provides a variety of search fields (e.g., title, content, provider, generator, access restriction, year of generation, type, format, and access method) and search

operators, and allows users to perform facet-based refinement of their search results. Second, data-driven analysis environments provide infrastructure for general or professional users to perform artificial intelligence-driven or deep learning-based research data analyses in a cloud environment. Functions for general users include the uploading of data or analytical models and the ability to generate, edit and execute an analytical workflow in a drag & drop manner. Functions for professional users include the provision of an infrastructure as a service (IaaS) environment, which allows researchers to configure their own analysis environments. Third, research data repository hosting services are especially helpful for institutions and research groups that want to manage research data more effectively but have difficulties due to the limited computing resources, a lack of a management system, or a shortage of human experts for system management. Fourth, to facilitate convergence and collaboration among researchers or research groups, research community services support the creation of thematic research communities, communication among researchers, and the sharing and utilization of research data. Moreover, linked to various data services, including NTIS, OpenAIRE (https://www.openaire.eu), Research Data Australia (https://researchdata.edu.au), and NII (https://www.nii.ac.jp), it provides one-stop discovery services for research data from national R&D projects.

Furthermore, we are attempting to establish a tiered research data sharing and utilization ecosystem at the national level, as shown in **Figure 12**. With the national research data platform, disciplinary data centers will be built hierarchically and research data from national R&D projects will be curated and utilized in the relevant disciplinary data centers. Six pilot projects focusing on bioinformatics, materials, large research equipment, and AI are being conducted to forge reusable data, and disciplinary sharing practices.

Institutional data repository

We are also developing an institutional data repository (**Figure 13**) for government-funded research institutes. It adopts Docker-based microservice architecture for scalability to accommodate domain-specific and institution-specific needs. On the basis of iRods (https://irods.org), it offers not only reliable data storage

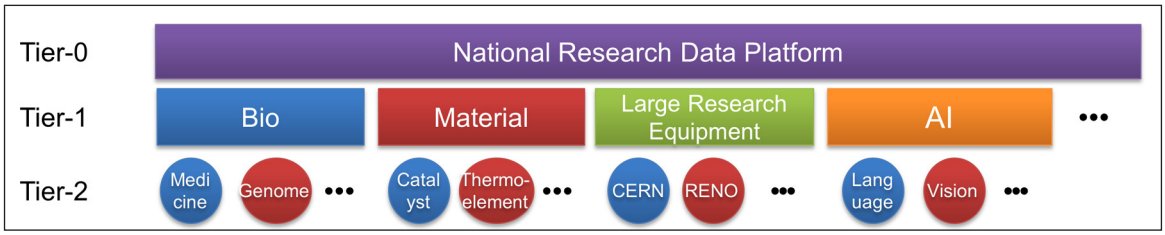

Figure 12: A tiered research data sharing & utilization ecosystem.

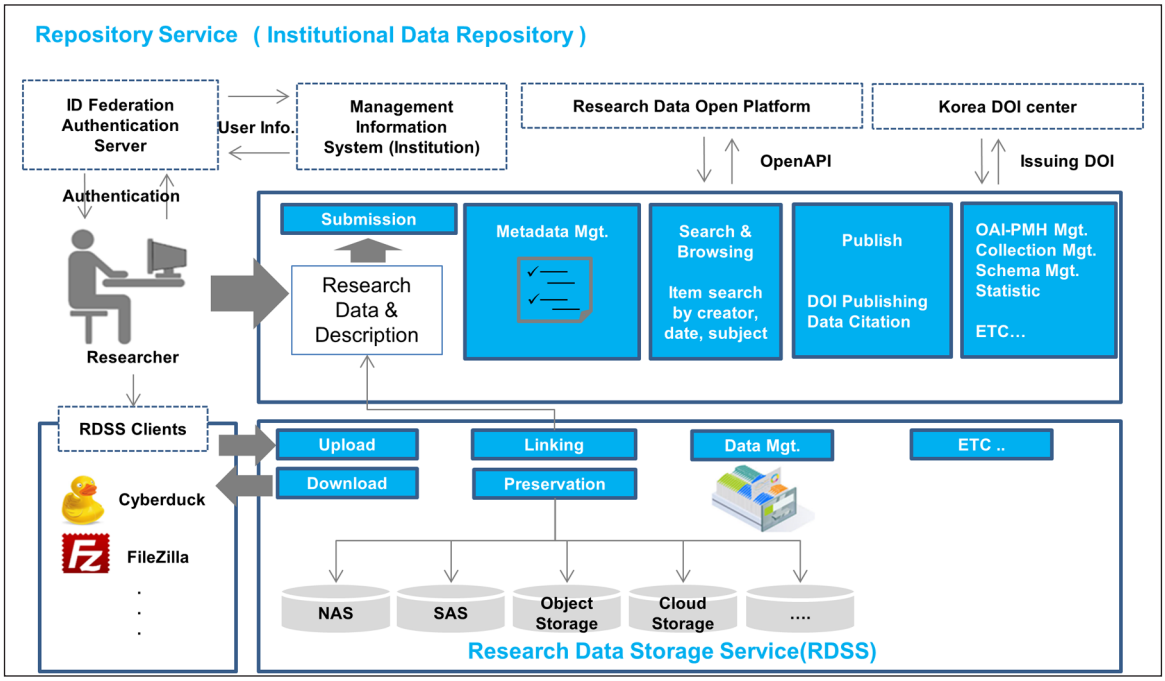

Figure 13: Institutional data repository for government-funded research institutes.

and preservation but also a private data storage service for archiving and collaboration during the research process. Research data is organized according to the collection-dataset-file hierarchy, and a standardized metadata management system allows for multiple customizable schemas for various types of research data. This repository will also provide federated ID authentication, including ORCID (https://orcid.org) and KAFE (https://www.kafe.or.kr). We plan to incorporate DMP creation and management tools into the repository in order to support the entire research data life cycle from planning to sharing.

Conclusion

In Korea, research data of various types and sizes are being created in national R&D projects, but they are usually stored and managed separately at the personal or laboratory level and are shared only upon a personal request. This occurs because the construction of data is not considered a major research output, and the management and sharing of data is a burden to researchers due to additional management tasks and responsibility issues. While a regulation that mandated a DMP in national R&D projects was put into effect recently, it remains urgent to devise follow-up policies and to develop the infrastructure for sharing research data. To promote the sharing and reuse of research data, research data itself should be recognized as a first-class research object. A national data management and utilization system, including national data discovery services and disciplinary data centers, represents a dire need for researchers. It would enable them easily to share and reuse their research data. Among various stakeholders, there are wide discrepancies in their points of view regarding the sharing and reuse of research data. Therefore, it is necessary to make consistent efforts to reach a consensus on the issue of open research data.

It is becoming increasingly important to increase the reusability and transparency of research through nationwide research data sharing and utilization efforts. We hope to create an open science ecosystem by creating, managing, and utilizing research data in a FAIR way. Much time will be needed to implement this system, but once complete it will provide a firm foundation for the next generation of scientific research.

Additional File

The additional file for this article can be found as follows:

- **Appendix.** RDM survey questionnaire. DOI: https://doi.org/10.5334/dsj-2020-029.s1

Funding Information

This research was supported by Korea Institute of Science and Technology Information (KISTI).

Competing Interests

The authors have no competing interests to declare.

References

Baker, M. 2016. Is There a Reproducibility Crisis? *Nature*, 533(7604): 452–454. DOI: https://doi.org/10.1038/533452a

Barsky, E. 2015. Research Data Management Survey: Science and Engineering. *R. Library Staff Publications and Research*, University of British Columbia. DOI: https://doi.org/10.14288/1.0348069

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

Begley, C and **Ellis, L.** 2012. Raise standards for preclinical cancer research. *Nature*, 483(7391): 531–533. DOI: https://doi.org/10.1038/483531a

Economist. 2016. Excel errors and science papers. 7 Sep.

Freedman, L, Cockburn, I and **Simcoe, T.** 2015. The Economics of Reproducibility in Preclinical Research. *PLOS Biology*, 13(6). DOI: https://doi.org/10.1371/journal.pbio.1002165

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

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

Kim, Y and **Yoon, A.** 2017. Scientists' data reuse behaviors: A multilevel analysis. *Journal of the Association for Information Science and Technology*, 68(12): 2709–2719. DOI: https://doi.org/10.1002/asi.23892

OECD. 2015. Making Open Science a Reality. *OECD Science, Technology and Industry Policy Papers.*

Open Research Data Forum. 2016. Concordat on Open Research Data. *Open Research Data Forum Report.*

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

Shearer, K and **Furtado, F.** 2017. COAR Survey of Research Data Management. *COAR Report*. Available at https://www.coar-repositories.org/files/COAR-RDM-Survey-Jan-2017.pdf.

Tenopir, C, et al. 2015. Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. *PLOS ONE*, 10(8). DOI: https://doi.org/10.1371/journal.pone.0134826

The Royal Society. 2012. Science as an open enterprise. *The Royal Society Science Policy Centre Report*.

Wilkinson, M, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3.

How to cite this article: Choi, M-S and Lee, S. 2020. Research Data Management Status of Science and Technology Research Institutes in Korea. *Data Science Journal,* 19: 29, pp. 1–11. DOI: https://doi.org/10.5334/dsj-2020-029

Submitted: 09 January 2020 **Accepted:** 08 July 2020 **Published:** 05 August 2020

]u[*Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS