Korean Case Report on Enhanced Access to Research Data

Version 1.2 (June 27, 2018)

Science and Technology Policy Institute (STEPI)
Eun Jung Shin, Ph.D., ejshin@stepi.re.kr

This report was prepared for the case study on the research data initiative led by the Science and Technology Policy Division of the OECD
Korean Case Report on Enhanced Access to Research Data

1. Overview

The Korean government, in particular the Ministry of Science and ICT (MSIT), has begun developing policy measures to enhance access to research data drawn from publicly-funded research projects since last October. The currently approved national strategy, the Strategy to Promote Sharing and Use of Research Data for Innovative Growth, proposes a more systematic approach with multiple policy measures. The Strategy is expected to greatly influence the way in which research data is produced, managed and shared among researchers in South Korea. This Korean case report will specify background information, key content and organizations, and expected outcomes of the Strategy.

Table 1. Overview of the Initiative

<table>
<thead>
<tr>
<th>Name of Initiative</th>
<th>Strategy to Promote Sharing and Use of Research Data for Innovative Growth (in South Korea)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objective</td>
<td>Aims to promote sharing and use of research data and eventually induce new discovery and innovation at the national level</td>
</tr>
<tr>
<td>Type</td>
<td>National Strategy, accompanied by policy measures (investment plans, legislative revisions, etc.)</td>
</tr>
<tr>
<td>Responsible Policy-making Body</td>
<td>-Lead Institute: Ministry of Science and ICT (MSIT)</td>
</tr>
<tr>
<td>Responsible Implementing Bodies</td>
<td>MSIT and Korea Institute of Science and Technology Information (KISTI), in close collaboration with other public research institutes</td>
</tr>
<tr>
<td>Target Audience</td>
<td>Korean researchers, small and medium-sized enterprises, and the public (scientific community, business, civil society, and the general public)</td>
</tr>
<tr>
<td>Total Duration of the Initiative</td>
<td>July 2017 – January 2018 (planning), January 2018 – current (implementation)</td>
</tr>
<tr>
<td>Sectoral Focus</td>
<td>Research &amp; Development</td>
</tr>
<tr>
<td>Type of Data Concerned</td>
<td>Research data obtained from government-funded national R&amp;D projects</td>
</tr>
<tr>
<td>International References</td>
<td>OECD, US, UK, EU and Australia</td>
</tr>
<tr>
<td>Expected Results</td>
<td>The Strategy presented is expected to increase awareness of the value of research data and facilitate scientific and socio-economic efforts to create tangible outcomes from shared research data. Eventually, the Strategy will promote sharing and use of research data across different areas of research and induce data-driven innovation and innovative growth.</td>
</tr>
</tbody>
</table>

2. Background Information, Rationale & Key Drivers of the Initiative

2.1. Policy Context Prior to the Initiative

(1) Development of Open Government Data Initiative & Open Access Policy on Public Research Outputs

- South Korea is a leading country in terms of open government data initiatives (OECD 2017, Government at a Glance); the Korean government has proactively invested in digital government infrastructure and e-government services since the early 2000s. In 2013, it introduced additional measures to increase access to and utilization of government data as part of the Government 3.0 Initiative (NIA
2013). Government agencies are increasingly involved in data tasks, identifying disclosable government data and providing it to the public in a more usable format, often prior to public requests for such data. As a result, South Korea has consecutively ranked at the top in the OECD OURdata Index on Open Government Data 2015/2017 (OECD website on Open Government Data)\(^1\).

In line with this open government initiative, the Ministry of Science and ICT (MSIT) established a national information system, the National Science and Technology Information Service (NTIS). Since 2006, NTIS has basically provided the public with core information on national R&D projects funded by the Korean government, which covers basic information on government R&D projects, who conducts them, and the produced outputs (MSIT website on NTIS)\(^2\). In addition, MSIT has developed policy tools to disclose and disseminate the range of these research outputs since the early 2000s. As a typical example, the *Regulations on the Management of National Research and Development* specifies researchers’ obligation to disclose and deposit the research outputs of national R&D projects to designated open access repositories and make them available to the public (National Law Information Center website)\(^3\). These *Regulations* set the basic rules on open access to publicly-funded research outputs.

- **(2) Increased Demands for Data in the Digital Economy**
  
- With widespread diffusion of digital technology, the capacity to create and utilize data has increased exponentially, which is expected to foster knowledge creation, economic growth and societal well-being in this age of the digital economy (OECD 2015, OECD 2016). Anticipating an increasing role of data in the near future, the Korean government established a Public-Private Joint Taskforce on Big Data in 2016. As government agencies, the Ministry of the Interior and Safety (MOIS), MSIT, and Statistics Korea are participating in the taskforce, and private-sector managers from Internet service providers, data providers and data-driven entrepreneurs were invited as well. Since then, the taskforce has discussed national strategies to promote a “big-data industry”. During the Moon Jae-in administration (2017–), the taskforce is collaborating with a new committee, the Presidential Committee on the Fourth Industrial Revolution. Early this year, the Committee issued *The National Plan for the Fourth Industrial Revolution*, which includes investment plans to enhance the data infrastructure and services and eventually promote data-driven innovation (Presidential Committee on the Fourth Industrial Revolution 2018). Especially regarding research data, *The National Plan for the Fourth Industrial Revolution* suggests providing more systematic support for sharing data and its use by researchers (Presidential Committee on the Fourth Industrial Revolution 2018).

---


(3) Policy Gaps & Challenges

- **Limited Access to and Use of Research Data**: With the existence of open government data initiatives and open access policy on publicly-funded research outputs, some research data is released as government data while some is disclosed as public research outputs. Nevertheless, research data is not a major policy target for either initiative. The open government data initiative mostly deals with administrative data generated by government services, while public research outputs mostly concern publications and patents. Software, biological materials, chemical compounds, and some technical information are now considered research outputs, but a wide range of diverse research data is not fully acknowledged as such yet. Furthermore, reward systems for the new types of research outputs are not well established (Shin, et al. 2016). It is challenging to motivate researchers to share their research data and materials with others. People outside research teams still feel that not much research data is available or usable.

- **Unmet Needs and Demands for Research Data**: On the other side of the supply-side shortage, is an expected increase in demand for research data. In contemporary science, research data is required not only to validate theory-based hypotheses but also to generate new hypotheses and theories (OECD 2015). A large amount of research data generated by research equipment and facilities has become another source of knowledge, which often paves the way for new theory-building (OECD 2017, OECD 2016). Advanced data mining and simulation techniques also allow researchers to leverage research data to drive discovery and innovation (OECD 2015, OECD 2016). Nowadays, research data is considered not only as a source of new scientific discovery but also as a source of national competitiveness and innovative growth in the Fourth Industrial Revolution (Presidential Committee on the Fourth Industrial Revolution 2018, Steering Committee of the National Science and Technology Council 2018). Although researcher demands for data storage and analysis are expected to surge, the current domestic infrastructure and human resources remain insufficient (Steering Committee of the National Science and Technology Council 2018). In addition, more effort is needed to support data-sharing practices and induce collective outcomes (Steering Committee of the National Science and Technology Council 2018).

- As a way to fill the gap between expected high demand for research data and its limited supply, the Korean government adopted early this year a national strategy developed in the second half of 2017, known as the “Strategy to Promote Sharing and Use of Research Data for Innovative Growth” (Steering Committee of the National Science and Technology Council 2018).

---

4 The Fourth Industrial Revolution is defined as a hyper connectivity-based intelligent technology revolution triggered by the development of artificial intelligence (AI), big data, and other digital technologies (Presidential Committee on the Fourth Industrial Revolution 2018, p. 12). The Presidential Committee on the Fourth Industrial Revolution also deals with data policies to prepare for the digital economy and society while emphasizing revolutionary and transformative nature of digital transition.
2.2. Rationale & Key Drivers of the Initiative

(1) Underlying Motives
- The *Strategy to Promote Sharing and Use of Research Data for Innovative Growth* envisions research data as a source of intellectual property as well as an essential part of big data (Steering Committee of the National Science and Technology Council 2018). Anticipating a dominant role for research data in scientific discovery and socio-economic development in the near future, the *Strategy* was devised as a means of leveraging it to enhance national competitiveness.

(2) Objectives as Defined
- As the title says, the main objectives of the *Strategy* is to promote sharing and use of research data at the national level (Steering Committee of the National Science and Technology Council 2018).

(3) Bodies Involved in Designing & Initiating the *Strategy*
- The *Strategy* was initiated by Minister of Science and ICT Young Min You as he entered office in mid-2017. He began by declaring the importance of research data and data-sharing in the digital economy and created a related entity known as the “Collected and Renewed” taskforce, after which government officials from the Ministry of Science and ICT (MSIT), policy experts, researchers, data scientists, private-sector managers, and lawyers were invited to a series of taskforce meetings. Between July and December 2017, the taskforce met 26 times and outlined a draft of the *Strategy*.

(4) International References and Good Practices as Drivers
- International policy trends on open science, research data, and digital science ignited domestic policy discussion on the *Strategy*. Widespread adoption of open science and open data policies in western countries was reviewed and introduced at the beginning of domestic discussions.
- The OECD’s 2016 working paper, which reviewed open science practices of member countries, was introduced in the final version of the *Strategy* (Steering Committee of the National Science and Technology Council 2018).
- In addition, the *Strategy* refers to current public policies of the US, the UK, the EU and Australia regarding open access to government data and research outputs, adoption of research data management tools, and support for data-intensive science and digital transition (Steering Committee of the National Science and Technology Council 2018). The fact that these countries and regions with advanced scientific capacities are involved in making policy on and investing in research data urged Korean decision-makers to develop a national strategy in this area as well. At the same time, the Korean government assured the public that other nations had proven the effectiveness of similar strategies.
- In a practical sense, the taskforce used international experience as benchmarks when deciding a definition of research data, ways it could be effectively shared and utilized, and the policy measures that were needed. For instance, the *Strategy* defines research data as “recorded factual material/data that is obtained from observations, experiments, surveys, analysis and other research activities of government-funded research projects and deemed as objective and necessary to validate research findings,” referring to the definition of research data in US
OMB Circular A-110 (Steering Committee of the National Science and Technology Council 2018). In addition, the Strategy introduced the US OSTP Memorandum (2013) as one example of policy requesting data management plans in national R&D projects (Steering Committee of the National Science and Technology Council 2018).

In sum, international references and good practices, including those discovered by the OECD, provided important guidance for the Strategy and certainly stimulated domestic adoption of it.

3. Main Content & Structures of the Initiative

3.1. Main Content/Tasks

(1) Objectives/Goals
- The main objective of the Strategy is to promote sharing and use of research data at the national level.

(2) Scope of Initiative (research data covered)
- The Strategy covers research data obtained from national R&D projects--that is, R&D projects funded by the Korean government. Research data itself can be final outputs as well as throughputs, and basically refers to recorded factual data necessary to validate research findings. However, the scope may vary across different areas of research.

(3) Key Policy Aims
- To promote sharing and use of research data, the Strategy proposes to achieve at least five policy aims over the next five years: (1) support establishment of a research data center in each field of research, and help it grow within the research community, (2) establish a national research data platform that connects field-oriented data centers to one another, (3) provide education and training for data scientists/engineers to enhance their data skills and expertise, (4) develop a legal basis for sharing and use of research data and adoption of data management plans, and (5) promote innovation and commercialization activities inspired from this shared research data (Steering Committee of the National Science and Technology Council 2018).

(4) Strategies with Policy Measures
- In order to achieve its objectives, the Strategy specifies five core sub-strategies: (1) issue some basic principles on data sharing and use, (2) promote community- and capacity-building processes in each area of research, (3) reduce researcher burden through a digital infrastructure and additional support, (4) apply data-sharing strategies to data-intensive research first before expanding to other fields on an incremental basis, and (5) conduct pilot projects and demonstrate existing needs and the potential value of data sharing in a relatively short time period (Steering Committee of the National Science and Technology Council 2018).
3.2. Governance, Structures & Process

(1) Planning
- **Lead Ministry or Institution**: The Ministry of Science and ICT (MSIT) was the lead institution that initiated the whole process. Under the auspices of a new minister, the High-Performance Computing (HPC) team set up a taskforce meeting on research data to develop a national strategy, composed of one overarching group and five specialized subgroups. Personnel from MSIT joined a diverse range of stakeholders such as researchers and data scientists from universities and government research institutes, private-sector managers, policy experts and lawyers.

- The taskforce began in July 2017 and completed its agenda in December over the course of 26 meetings. Based on these tasks, MSIT’s HPC team developed a draft of the *Strategy to Promote Sharing and Use of Research Data*. Then, the ministry held a public hearing and incorporated feedback from the public into the final draft of the *Strategy*. After review by the Minister of Science and ICT once again, the *Strategy* was submitted to the Steering Committee of the National Science and Technology Council (NSTC).

- **Relevant Bodies/Organizations/Groups**: NSTC is a high-level inter-ministerial coordination committee, and is tasked with developing a comprehensive national plan on science, technology and innovation and making adjustments to R&D programs and budgets proposed by each ministry. The Steering Committee of the National Science and Technology Council reviews specific strategies and plans associated with the national plan on a regular basis. The *Strategy on Sharing and Use of Research Data* was reviewed and approved by this steering committee in January 2018, and has been implemented mainly by MSIT since then.

- Figure 1. Governance of the Initiative during the Planning Stage

(2) Implementation
- **Lead Ministry or Institution**: MSIT governs implementation of the *Strategy*, by further developing action plans with relevant bodies and organizations and making them reality. The Basic Research on Information & Electronics Team (previously the HPC Team) within MSIT is the specific entity presently leading
implementation, while actively collaborating with divisions within MSIT, government funding agencies, government research institutes, digital research infrastructure managers and other stakeholders.

- Relevant Bodies/Organizations/Groups: In addition to the related divisions within MSIT, government funding agencies such as the National Research Foundation are considering developing their own policies on research data in line with the Strategy, and applying for contract-based government R&D projects. In particular, large-scale government R&D projects and government research centers will be tasked with developing research data policy (including data management plans) earlier than others. The National Research Council of Science and Technology (NST), an umbrella organization of government research institutes, is also developing guidelines to assist government research institutes with management of their research data.

Figure 2. Governance of the Initiative during the Implementation Stage

Source: Revised from the Steering Committee of the National Science and Technology Council (Jan. 19, 2018), p. 6.

- Pilot Projects/Data Centers: At the same time, pilot projects on research data are now being implemented. Although the Strategy proposes the establishment of a National Research Data Platform (NRDP) and field-oriented data center hubs, a national investment plan would take significant time and require significant resources to be reviewed and approved by the relevant administrative bodies. Even before establishment of such an NRDP and data center hubs, pilot projects have been designed to facilitate data-sharing in small areas of research and induce innovative research outcomes in a relatively short time as a way to motivate other research groups to participate in subsequent projects. This year, pilot projects are being led by several research groups belonging to government research institutes that have already invested in data-intensive research through their own research
data centers. This year, at least six such centers across four different fields of research are involved.

(3) Status of Implementation
- The *Strategy to Promote Sharing and Use of Research Data* presented five key policy aims to be accomplished in five years. The details of their status of implementation can be seen in Table 2.

<table>
<thead>
<tr>
<th>Key Policy Aims</th>
<th>2018</th>
<th>2019-2022</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Support establishment of a research data center in each field of research, and help it grow within the research community</td>
<td>Pilot projects in data-intensive research</td>
<td>Scope of projects to expand</td>
</tr>
<tr>
<td>(2) Establish a national research data platform that connects field-oriented data centers to each other</td>
<td>Preparations for budget review</td>
<td>New projects launched</td>
</tr>
<tr>
<td>(3) Provide education and training for data scientists/engineers to enhance their data skills and expertise</td>
<td>Launched small-scale program</td>
<td>Programs to scale up</td>
</tr>
<tr>
<td>(4) Develop a legal basis for sharing and use of research data and adoption of data management plans</td>
<td>Began revision of major legislation</td>
<td>Revise relevant regulations and guidelines</td>
</tr>
<tr>
<td>(5) Promote innovation and commercialization activities inspired from this shared research data</td>
<td>Provide assistance linked to existing programs on commercialization of research outputs and the R&amp;D service industry</td>
<td>Run specialized centers for utilization and commercialization of research data</td>
</tr>
</tbody>
</table>

4. Status of Implementation in Four Areas of Research

Pilot projects on research data aim to support existing extraordinary data-sharing systems to bring about tangible research outcomes. Therefore, this year’s pilot projects are mostly run by government research centers that adopted data-sharing systems years ago and have created meaningful research outcomes. From the perspective of a research center, a pilot project is supplementary to its mainstream data-sharing project. Therefore, it is necessary to understand the overall activities of these mainstream projects in addition as they compare to the activities in the pilot project. This section looks at six research centers participating in a pilot project this year while running their own mainstream projects. The six projects, which lie in four different areas of research, are presented below.

4.1. Pilot Projects on Bio-research Data

(1) Overall Plan
The Strategy proposes development of a research data center hub that can support field-specific needs and demands in bio-research, and connect it with individual data centers in small areas of research. As a short-term task, pilot projects were proposed to create meaningful research outcomes from pre-existing data-sharing practices in small fields of research. This year saw the beginning of two pilot projects: one in the Korean Bio-information Center (KOBIC) and the other in the Medicinal Biochemistry Division of the Korean Research Institute of Chemical Technology (KRICT).

(2) CASE 1: Pilot Project on Genome Data in KOBIC
- **Mainstream Project:** The Korean Bio-information Center (KOBIC) is running its own project (2014-2022) involving management of genome data drawn from the Post-Genome Project. According to project regulations, participating research organizations are required to place genome data produced during the project to a designated KOBIC repository after project completion. KOBIC develops one-stop portal services on genome data and helps with public access to the collected genome data. In addition, KOBIC provides cloud analytics for large-scale genome data, a so-called “Bio-Express” service.
- **Pilot Project:** A pilot project on research data aims to facilitate utilization of genome data collected in KOBIC. This year, KOBIC held a Genome Analytics Competition, and asked the public to propose ways to improve the current analytics of Bio-Express (Steering Committee of the National Science and Technology Council 2018).
- **Research Data and Users:** So far, 105TB data have been drawn from 131 projects and been made available to the public. Major users of genome data provided by KOBIC include Korean researchers working at universities, government research institutes, companies, and research hospitals. Teachers, lecturers, and trainers also use this data for educational purposes.
- **International References:** The American Society of Human Genetics (ASHG), the European Society of Human Genetics (ESHG), and the International Society for Computational Biology (ISMB)
- **Data Governance/Trust:** Human genome data is subject to the Personal Information Protection Act in South Korea. In order to protect individual privacy, human genome data must be protected by encryption and transferred only after de-identification or anonymization of personal information.
- **Data Quality/Standards:** Collected genome data is rather fragmented with meta-data, which hinders subsequent sharing and use. Data quality relies mainly on data providers, whereas KOBIC managers only check for a basic level. More effort is needed to improve data quality according to global standards.
- **Data Ownership/Rewards and Incentives:** KOBIC does not have its own policy on data ownership but complies with national regulations on R&D outputs. Extra rewards or incentives do not exist for those who deposit quality data.
- **Business Models:** KOBIC’s genome data repository is government-funded.
- **Data Skills and Human Resources:** Professional skills and staff are needed. Given the rapid development of genome analytics, continuous education/training of managers and technicians is necessary to keep them up-to-date with qualifying skills and knowledge.

(3) CASE 2: Pilot Project on New Drug Discovery Based on the KRICT Chemical Library
- **Mainstream Project:** The Medicinal Biochemistry Division of the Korean Research Institute of Chemical Technology (KRICT) has run a chemical library with government funding since the 2000s. More than 550,000 chemical compounds and millions of experiment logs on 760 drug targets have been collected from government research institutes, universities, and companies. These compounds and data are available to users only upon agreement/informed consent from the data producers. Public access is possible once data providers allow it, and this amounts to a million test logs on 130 drug targets. KRICT also developed a Chemical Information Management System (CIMS) and has improved the Library’s information management flows. CIMS allows users to instantly check degree of openness (e.g. Closed, Exclusive to project members, Exclusive to institutional members, and Open to the public) and manage chemical information in a systematic and appropriate manner.

- **Pilot Project:** A pilot project (2018-2019) aims to promote discovery of new drugs by advancing a database of compound bioactivities and toxicities in the chemical library and applying AI technology to drug discovery (Steering Committee of the National Science and Technology Council 2018).

- **Data Governance/Trust:** Establishment of fair agreement systems is crucial, especially when chemical information is expected to generate profit from intellectual property and commercial rights. The chemical library is basically run through bilateral agreement systems between data producers and users; it also allows data producers to decide the degree of openness in terms of what is opened and who has access.

- **Data Quality:** Chemical library managers constantly check and control the purity and quality of chemical compounds and replace poor-quality compounds with new ones. Library managers also verify the quality of chemical data by replicating part of the experimental logs. Nonetheless, it is desirable to develop more user-friendly electronic laboratory notebooks and encourage data providers to use them to improve the reliability of experiment logs and data.

- **Data Ownership/Rewards and Incentives:** Although there exists no explicit policy on ownership of publicly-disclosed data, Library managers encourage users and providers to collaborate with each other and share the benefits from their research based on their contributions.

- **Business Models:** KRICT’s chemical library runs mostly on institutional funding and contract-based project funding from the Korean government, without significant input from the private sector.

- **Data Skills and Human Resources:** Professional skills and staff are needed. KRICT is planning to hire personnel specialized in data management.

### 4.2. Pilot Project on Future (Chemical) Material Data

(1) **Overall Plan on Future (Chemical) Materials Data**

The *Strategy* recommends the establishment of research data centers in each small field of research by granting diverse scientific communities to maintain their data-sharing systems that are aligned with the attributes of material research data. Field-oriented data centers include those on thermo-electronic materials and catalysts. At the same time, the *Strategy* calls for continuous efforts to curate and standardize research data across different fields of research, and eventually improve accessibility and usability of the data. A designated and specialized data center hub is proposed to
promote standardization and curation of future material data. Operations are expected to be in close cooperation with the existing data center on industrial materials: the Materials Center.

(2) CASE 3: Pilot Project on Thermo-electronic Materials in KRICT
- **Mainstream Project:** The Korean Research Institute of Chemical Technology (KRICT) is conducting a research project (2017-2022) to develop new quality thermo-electronic materials utilizing big data and machine-learning analytics. In addition to KRICT research teams, research teams from Seoul National University (SNU) and the National Institute of Mathematical Science (NIMS) are also participating. The research teams share their research data and analytics, which include unpublished research data generated by experimental failures. The project teams aim to file patents for the newly developed materials in the future.
- **Pilot Project:** A pilot project facilitates data-sharing and research collaboration through development of a web-based data platform on thermo-electronic materials and further developing machine learning analytics.
- **International References:** The National Institute of Standards and Technology (NIST) in the US runs a website on materials data based on the Materials Genome Initiative, and has recently developed the Materials Data Curation System. In Japan, the Materials Research by Information Integration Initiative (MI2I) was launched in 2015.
- **Data Governance/ownership:** Researchers must be granted ownership and intellectual property rights over their materials data. Therefore, the project teams consider opening access to materials data only after related publications and patents are released. It would be desirable to allow data providers to decide when and how their data is shared. Nevertheless, no explicit policy exists on these matters yet.
- **Data Rewards and Incentives:** Presently, there are no rewards or incentives for data providers. Unique identifiers (e.g. digital object identifiers) need to be developed which can be attached to original data and used in data citation.
- **Data Quality:** It is necessary to standardize methods for collecting and analyzing materials data across different fields.
- **Business Models:** KRICT projects are run with contract-based government project funding.
- **Data Skills and Human Resources:** Data scientists need to interact with researchers with domain knowledge and learn more about materials data throughout this kind of R&D project.

(3) CASE 4: Pilot Project on Catalyst Data in KIST
- **Mainstream Project:** The Korean Institute of Science and Technology (KIST) is one of the largest and oldest government research institutes in South Korea. KIST has recently invested in transforming its digital research infrastructure institute-wide (KIST R&D Informatics, or “KiRI”) for more data-intensive research. KIST members defined the basic concepts and functions of KiRI last year and are currently developing prototype service portals which will be available to KIST members in the middle of this year. KiRI is basically composed of two parts: (1) KiRI Note, which allows members to use electronic lab notebooks in daily life, and (2) KiRI Platform, which provides to members research data and analytical tools developed by their colleagues as well as external sources. As part of KiRI,
KIST is also planning to develop an R&D Data Bank to assist KIST members with systematic management of their research data.

- **Pilot Project**: A pilot project (2017-2018) facilitates the establishment of data-sharing systems (R&D Data Bank) in the field of catalyst research while reaffirming KIST’s investment plan in KiRI.

- **Data Governance/ownership**: KiRI is basically an Intranet shared with KIST members. KIST intends to grant ownership of research data to the data creators and allow them to withhold such data until they achieve their pursued objectives.

- **Data Rewards and Incentives**: Rewards or incentives for data-sharing in KiRI have not been considered yet.

- **Business Models**: KIST projects are run with government institutional funding.

- **Data Skills and Human Resources**: Researchers/experts with domain knowledge in each discipline need to interact with data scientists and engineers for appropriate management of field-specific data.

### 4.3. Pilot Project on Research Data from Large Research Facilities

(1) **Overall Plan on Research Data from Large Research Facilities**

The *Strategy* proposes promoting sharing of in-demand research data produced by large research facilities (e.g. data from electron microscope and particle accelerator). It also suggests expanding the scope of data sharing practices into a field of significance and synergy with information and communication technologies (e.g. astronomy, space, and nuclear data). Lastly, the *Strategy* recommends ensuring that quality research data collected from large research facilities (e.g. reference data) is accessible and reusable.

(2) **CASE 5: Pilot Project with Large Research Facilities**

- **Mainstream Project**: The Korea Institute of Science and Technology Information (KISTI) hosted the Global Science Experimental Data Hub Center (GSDC) in 2010 to support global data-intensive research that requires a high-performance computing infrastructure. A typical example of the assistance provided by GSDC is international research collaboration with CERN. In addition, Korean researchers are participating in global experiments led by LIGO (Laser Interferometer Gravitational-Wave Observatory) and KEK (Japanese High Energy Accelerator Research Organization) via a digital infrastructure provided by GSDC.

- **Pilot Project**: With high-performance computing systems, GSDC aims to retrieve research data generated by two different domestic research facilities (the Pohang Particle Accelerator and the KBSI High Voltage Electron Microscope), and provide it to researchers interested in integrative structural biology with analytical tools.

- **Data Governance/ownership**: Research data collected via a pilot project do not include personal information, negating the need for concern about privacy issues. Those who produce research data are considered to hold ownership over it, although explicit policy on that matter has not been developed. Collected data is only available to project teams participating in pilot project experiments at the moment.

- **Data Quality**: Participating research teams and the research community (Korean Society for Structural Biology) are responsible for reviewing the quality of
research data. GSDC is developing a standard data management plan (DMP) in consultation with the Korean Society for Structural Biology. In addition, GSDC is assisting managers of two large research facilities in developing facility data management plans.

- **Business Models:** GSDC projects are run with contract-based government project funding.

- **Data Skills and Human Resources:** It is desirable to develop training programs on research data management through cooperation with researchers with domain knowledge, research facility managers, and data scientists and engineers.

### 4.4. Pilot Project on Artificial Intelligence Data

(1) **Overall Plan on Research Data for Artificial Intelligence Research**

The *Strategy* proposes expanding the scope of data collected for the development of artificial intelligence (AI) technology. It also suggests establishing a data center hub for AI data and providing the public with collected research data through it. The data center hub is expected to promote collaboration among researchers, research agencies, start-ups, and other firms and eventually function as machine-learning testbeds.

(2) **CASE 6: Pilot Project on Research Data for AI Research**

- **Mainstream Project:** The National Information Agency (NIA) has developed a machine learning database for the advancement of AI. A majority of the data collected is publicly available government data, including on public transportation, patents and laws, as well as images and language data. In addition to investing in data, the NIA is developing AI algorithms/analytics and a supercomputing infrastructure. After integrating AI data, AI analytics, and supercomputing work environments, the NIA launched the AI Open Innovation Hub early this year.

- **Pilot Project:** The NIA is expanding the scope of research data collected and collaborating with government research institutes that produce useful data for AI development (e.g. ETRI, KETI)

- **Data Governance/ownership:** The AI Open Innovation Hub functions as a key platform for interaction between providers and users of research data, as the data available through it is basically open to the public. Data providers have the right to decide to open up their data or not, but when they decide to withhold data, the reason needs to be articulated. Rules and guidelines are needed on research ethics, protection of privacy, and security.

- **Data Quality:** NIA project team is collaborating with the Telecommunications Technology Association (in Korea) in order to develop standard procedures for AI data processing. Since international standards (ISO/IEC JTC 1/SC 42) are set up, more efforts need to be made to develop standard procedures and schema on AI data.

- **Data Rewards and Incentives:** There exist no direct rewards for data providers. Increasing next-year budgets for the institutes providing a large amount of quality AI data this year would be a good move.

- **Business Models:** NIA projects have been run with contract-based government project funding so far. However, public agencies are not the only entities interested in AI data: private enterprises also seek access, meaning that business
models for AI data can be created. Domestic AI data centers based on public-private partnerships are also needed.

5. Evaluation, Lessons & Challenges

5.1. Monitoring & Early-Stage Evaluation

(1) Measures to sustain goal-alignment and transparency of the initiative/ the role of evaluations in improving initiative design
The Strategy to Promote Sharing and Use of Research Data has been further developed into action plans by related divisions of the MSIT. The Minister of Science and ICT convened a couple of meetings to monitor the progress so far, while the Basic Research on Information and Electronics Team (previously the HPC team) within MSIT is monitoring policy actions taken for the Strategy on a daily basis. The Strategy as reviewed by the NSTC is publicly available, which guides participants to align their projects/programs with the goals specified within. In implementation, MSIT personnel have formal and informal meetings with participants and stakeholders to refine the goals and action plans.

(2) Criteria and key indicators for evaluating the initiative
Most of the projects initiated by the Strategy are government funded, and are evaluated according to the scientific and socio-economic outcomes, just like other government-funded projects. Since the Strategy highlights the achievement of tangible outcomes from sharing and use of research data, performance management is an essential part of project management. However, there exists no single performance indicator/target imposed on the projects, because the Strategy embraces a diversity of projects and plans across different fields of research.

(3) Impact assessment on end users/limitations and challenges in monitoring and evaluation
No impact assessment on end users is planned at the national level yet, as the Strategy was initiated only a year ago, making it too soon to identify any independent effects.

5.2. Main Issues Covered/Not Covered in the Initiative

(1) Data governance for trust – addressing privacy, confidentiality, quality and ethical issues
The Strategy proposes running a national research data platform which would connect individual data centers to each other. However, it does not specify roles and responsibilities of each entity, which allows data centers and related organizations to develop their own plans suitable to their individual situations. Data centers participating in pilot projects this year substantially vary in terms of the data concerned, the relationship between data providers and users, the roles and functions of data centers, and data repository openness to the public. Therefore, privacy, security, confidentiality and quality of data differ across data centers. A project on genome data cares more about protecting privacy (Case 1), whereas projects to develop particular technologies care more about intellectual property rights (Case 2 & 3). Although each project is subject to national laws on protecting privacy, intellectual property, national security, and other related issues, it is challenging to review all the ethical and legal
aspects of data sharing and use at data centers at the moment. Most centers participating in pilot projects do not have their own ethics review bodies yet.

(2) Discoverability/findability, machine readability and data standards

The Strategy emphasizes the scientific and socio-economic value of research data in the age of the digital economy and the Fourth Industrial Revolution. To realize the potential of research data, it needs to be discoverable, accessible, interoperable, usable and reusable. In particular, machine readability becomes increasingly important as research data is subject to big data analytics and AI technology (Cases 1-6). Open source software that helps with analysis of research data is also essential to data-sharing (Case 1 & 6). There still exists a gap between need and supply in data standards and quality management guidelines. To fill this gap, government funding agencies and research institutes are now considering adopting data management plans for major government projects in certain areas of research (e.g. data-intensive research). Although domain-specific guidelines and data standards increase over time, it would be meaningful for diverse research and data communities to develop more general standards and guidelines on data quality management at the international level.

At the same time, it is necessary to consider various situations in which research data is created and shared these days, when discussing accessibility and openness. Research data is not open by default in many cases. Some is exclusive to project members during a certain project period (Case 2 & 3), while some is open to a specific group of people on a conditional basis (Case 4 & 5) for security, IP issues, and other reasons. The existence of unavoidable or strategic conditions in which research data is not open by default needs to be acknowledged.

(3) Recognition and reward system for data authors

The Strategy seeks to boost intellectual property rights and socio-economic benefits from sharing research data between individuals. Even though no specific rules exist on how to divide these benefits or with whom at the national level, research teams involved in pilot projects have implicit rules to grant data providers a certain period of time to utilize their data first and let them decide to what extent their data will be disclosed. Nonetheless, guidelines on sharing the rewards/incentives created from data-sharing with data providers, data users, and data curators and managers remain vague. One data center requires bilateral agreements between data providers and users (Case 2), whereas most centers do not have any policy on data ownership, rewards or incentives. Unlike publications and patents, reward systems for research data are still under development. Only a couple of data centers are considering adoption of unique identifiers of data and promote data citations, given the complex and indefinite nature of dealing with research data. More effort is needed to institutionalize reward systems so that those who are involved in data creation, mining and curation can share in the reasonable benefits with data end-users.

(4) Business models for provision of open data

The Strategy proposes several government projects funded by the Korean government, and pilot projects this year are mostly government-funded. However, some areas, such as AI data, are receiving attention from private-sector entities, which opens up the possibility of developing public-private partnerships for open data provision.

(5) Building human capital and institutional capacities at public agencies to manage, create, curate and reuse data

The Strategy identifies the insufficiency of human capital and institutional capacities in data management and suggests the development of education/training programs for data scientists and engineers. This year, the Korea Data Agency is providing training programs to
potential providers and users of research data in line with the previous program on big data analytics. Educational programs are expected to increase as data centers specialize in diverse areas of research. Data center managers participating in the current pilot projects also call for the development of data scientists and engineers in each discipline with domain knowledge. The importance of human capital and institutional capacity in management and utilization needs to be acknowledged when discussing access to research data. Even though more and more data is now available and accessible, there will be no use of it if institution and individuals do not know how to properly gain access to or make use of it.

5.3. Lessons & Challenges

The Strategy revitalizes policy discussion on open science, data-intensive research, data-driven innovation, big data industry, and the Fourth Industrial Revolution. It is too early to assess its direct impact, but it has certainly wrought changes in government actions. A couple of related divisions within MSIT are preparing to revise policies and investment plans on research data, while government funding agencies and research institutes are doing the same on research data management. However, it is noteworthy that most researchers are still reluctant to share their research data with others as they remain concerned about the associated risks and burdens (An et al. 2017). Therefore, as suggested in the Strategy, it would be meaningful to begin with a research community that acknowledges the need for data-sharing and has the desire to share theirs with others to produce certain types of research outcomes. In addition, key investment plans specified in the Strategy (including establishment of the National Research Data Platform) are subject to a Preliminary Feasibility Review on (Major) R&D Projects. Like other major national projects, the investment plans for the Strategy will go through such a review which systematically assesses the scientific, economic and policy impacts of the investment plans and decides whether to implement them or not. Detailed actions plans for the Strategy may change in accordance with the results of this review.
An et al. (2017) conducted a national online survey of Korean researchers in an effort to understand current open science perceptions and practices. Approximately 10,000 researchers who were engaged at one time in government R&D projects and work at universities or government research institutes were invited to participate. The final respondents totaled 787 individuals: 473 from universities and 314 from government institutes. The survey ran for about two weeks in 2017, from January 4th to 19th. In addition to general background information on each individual, the survey investigated perceptions of and practices with open access, open data and open science at large. In the survey, only 36% of respondents said they had ever disclosed their data, while the majority had not. The reasons included concerns related to privacy, security, infringement of intellectual property, and misuse of data.

**Figure A. Percentage of Responding Researchers Who Had Disclosed Their Data & Underlying Reasons**

| Disclosed data associated with publications | 19.3 |
| Disclosed part of an entire dataset | 10.4 |
| Disclosed an entire project dataset | 6.8 |
| Did not disclose any data | 63.4 |

**WHY did you disclose your data?**
- To publicize research results and improve credibility
- To comply to sharing norms within scientific community
- Requested by funding agencies, publishers and etc.
- In expectation of socio-economic benefits

**WHY didn’t you disclose your data?**
- To protect privacy/security
- To protect for future publication
- To prevent misuse of data
- To avoid burden/responsibility associated with data disclosure
- No expectation of substantial benefits from data disclosure


References


The Steering Committee of the National Science and Technology Council (Jan. 19, 2018) Strategies to Promote Sharing and Use of Research Data for Innovative Growth, Ministry of Science and ICT.

U.S. OMB Circular A-110