

World Bank Group Korea Office
Innovation and Technology Note Series

JANUARY 2024, NOTE SERIES NUMBER 11

Enabling Data-Driven Innovation

Learning from Korea's Data Policies and Practices for Harnessing AI



Acknowledgements

This case study was made possible by the support of the Korea Digital Development Program (KoDi) of the Digital Development Global Practice at the World Bank. The drafting team authoring the report comprised Zaki B. Khoury (Senior Digital Development Specialist and Task Team Leader, Korea Digital Development Program); Yoon-seok Ko (Executive Principal, NIA); Seok-Jin Eom (Professor, Seoul National University); Keon Chul Park (Senior Researcher, Advanced Institute of Convergence Technology); Jung-Eun Park (Research Fellow, NIA); Bora Cho (Manager, NIA); Jisoo Lee (Consultant, World Bank); and Yulia Lesnichaya (Consultant, World Bank).

The authors are grateful to the Ministry of Science and ICT (MSIT) of the government of Korea for providing valuable inputs and information for this case study, as well as the valuable contribution received from the National Information Society Agency (NIA). The authors also appreciate the guidance and support received from Jason Allford, Special Representative to Korea at The World Bank; Mahesh Uttamchandani, Practice Manager, East Asia Pacific (EAP) region, Digital Development; and Malarvizhi Veerappan, Program Manager and Senior Data Scientist, Development Economics at the World Bank.

Special thanks go also to the following individuals for their valuable comments: Yong-jin Lee, Executive Director of NIA; Professor Tae-Woo Nam of the Sungkyunkwan University; Professor Kyung Ryul Park of the Korea Advanced Institute of Science and Technology (KAIST); Mark Williams; Jen JungEun Oh; as well as Oleg V. Petrov, Rong Chen, Jonathan Marskell; Sharmista Appaya, and Toni Kristian Eliasz of the World Bank Digital Development Global Practice. The authors are also thankful to Luba Vangelova and Sunny Kaplan for providing editorial guidance, the Korea Digital Development Program (KoDi) team, and The World Bank's Korea Office for overall support.

Rights and Permissions

The material in this work is subject to copyright. Because the World Bank encourages dissemination of its knowledge, this work may be reproduced, in whole or in part, for noncommercial purposes as long as full attribution to this work is given.

Any queries on rights and licenses, including subsidiary rights, should be addressed to World Bank Publications, The World Bank Group, 1818 H Street, NW, Washington, DC 20433, USA; fax: 202-522-2625; e-mail: pubrights@worldbank.org.

The Korea Office Innovation and Technology Note Series is intended to summarize Korea's good practices and key policy findings on topics related to innovation and technology. They are produced by the Korea Office of the World Bank. The views expressed here are those of the authors and do not necessarily reflect those of the World Bank. The notes are available at: <https://www.worldbank.org/en/country/korea>

Table of Contents

Acronyms and Abbreviations	8
Executive Summary	10
I. An Overview of Korea's Data Ecosystem	16
1.1 Background	16
1.2 Key Details of Korea's Data Policy	19
1.2.1 Development of Data Infrastructure and Platforms	19
1.2.2 Data Governance: Organizational Structure and Legal System	21
1.2.3 Data Services	25
1.3 Training and Fostering of Skilled Professionals	27
1.4 A Multi-Platform Data Ecosystem	29
1.4.1 Open Government Data Platform	29
1.4.2 Big Data Platforms	30
1.4.3 AI Training Datasets	31
II. Data Platforms	33
2.1 Open Government Data Platform	33
2.1.1 Core Assets: Data	33
2.1.2 Economical and Social Assets: User Service	34
2.1.3 Technical Assets: Hardware and Software Infrastructure	35
2.2 Big Data (Market) Platform	37
2.2.1 Core Assets: Data	37
2.2.2 Economical and Social Assets: User Service	39
2.2.3 Technical Assets: Hardware and Software Infrastructure	40
2.3 AI Hub	43
2.3.1 Core Assets: Data	43
2.3.2 Economical and Social Assets: User Service	43
2.3.3 Technical Assets: Hardware and Software Infrastructure	45
III. Analyzing Korea's Data Policies and Practices	47
3.1 Performance of the Data Policies	47
3.1.1 Results of Opening the Public Data	47
3.1.2 Performance of the Big Data Platforms	49
3.1.3 Performance of the AI Hub	53
3.2 Lessons Learned from Korea's Data Policies	55

3.2.1 Characteristics of Korea's Data Policy	55
3.2.2 Success Factors of Korea's Data Policy	56
3.3 Challenges in Korea's Data Practices	57
3.3.1 Data Collection and Areas of Generation	57
3.3.2 Aspects of Data Distribution and Trading	58
3.3.3 Aspects of Data Analysis and Utilization	58
3.4 Future Tasks to Enhance Korea's Data Practices	59
3.4.1 Establishing a Robust Data Governance Framework	59
3.4.2 Developing a Platform Focused on Consumers and Users	61
3.4.3 Transitioning from a Government-Led to a Market-Friendly Data Policy	62
IV. A Way Forward for Data Policy Making	64
4.1 Choosing the Right Time to Deploy the Policy	64
4.2 Adopting an Agile and Accurate Decision Making	66
4.3 Using the Collective Intelligence of “Crowd Workers”	66
4.4 Focusing on Quality, Quality, Quality	69
4.5 Prioritizing Privacy Sacrifice Quality, but Elevate Privacy	71
4.6 Building Institutional Capacity for Both Policy Making and Implementation	72
Bibliography	74

List of Figures

Figure ES.1: The Four Elements of Korea's Data Policy	10
Figure ES.2: Korea's Three Core Data Platforms	11
Figure 1: Changes in the Korean Government's ICT Policy	17
Figure 2: The Four Elements of Korea's Data Policy	19
Figure 3: Korea's Three Core Data Platforms	20
Figure 4: Korea's Digital New Deal: Data Dam	20
Figure 5: Korean Government Data Organizational Structure	21
Figure 6: Mid- and Long-Term Plan for AI Training Data Project	26
Figure 7: Conceptual Diagram for the AI Hub Service	27
Figure 8: Rate of Big Data-related Labor Shortage and Difficulties of Big Data-related Companies	27
Figure 9: Korea's Data Workforce Development System	28
Figure 10: Machine-Readable Data Stages	29
Figure 11: Disclosure Stages of Public Data	29
Figure 12: Conceptual Diagram of the Roles of Big Data Platforms and Centers	30
Figure 13: Data Development Process for AI Learning	31
Figure 14: Data Disclosure and Opening Procedures	32
Figure 15: Overview of the Public Data Portal	33
Figure 16: National Data Map	35
Figure 17: Configuration of the Public Data Portal System	36
Figure 18: Configuration of the Big Data Platforms and Centers	37
Figure 19: Configuration and Operating System of the Big Data Platform	39
Figure 20: Permission to Use by Data Type of the Healthcare-Related Big Data Platform	39
Figure 21: Expansion of the Utilization of Data Vouchers	40
Figure 22: Sharing and Interworking System of the Big Data Platforms	41
Figure 23: Object Identifier System	41
Figure 24: (Example) Development of a Platform and Center for the Field of Transportation	42
Figure 25: Data Sets Provided by the AI Hub	43
Figure 26: Configuration of the AI Hub Service	44
Figure 27: AI Hub's Corporate Support Service	44
Figure 28: Configuration of the AI Hub System	45

Figure 29: Cloud-Based Infrastructure Configuration of AI Hub	46
Figure 30: Concept of Object Storage	46
Figure 31: Cumulative Number of Opening Public Data (Unit: counts)	47
Figure 32: Number of Utilization of Public Data (Unit: counts)	47
Figure 33: Trend in the Proportion of Public Data in Open Format	48
Figure 34: Status and Trend Related to the Global Evaluation of Public Data	48
Figure 35: Status of Data Development for AI Learning	53
Figure 36: Number of AI Hub Visitors and Number of User Membership Sign-Ups	54
Figure 37: Performance in the Use of AI Hub's Key Services	54
Figure 38: Public Sector's Data Governance	59
Figure 39: Structure of the Data Ecosystem	61
Figure 40: Comparison of Korean and Japanese Interest in Artificial Intelligence, 2010-2021	64
Figure 41: Factors Hindering Companies' Adoption of AI and the Job Creation Effect (persons)	65
Figure 42: The Korean Government's Annual Investment Plan for AI Training Datasets	67
Figure 43: Rate of Use of the Internet by Age and Number of Users (%; thousand people; based on population ages 3 and older)	68
Figure 44: Korea's Data Verification Procedures	71

List of Tables

Table 1: Overview of the Big Data Master Plan and Big Data Service Activation	18
Table 2: Overview of Plan to Foster New Internet Industry	18
Table 3: Overview of the Data Industry Revitalization Strategy	19
Table 4: Key Contents of the Act on Promotion of the Provision and Use of Public Data	22
Table 5: Public Data Act of Korea: Provision and use of public data	23
Table 6: Key Contents of the Data Framework Act	24
Table 7: National Data Policy's Implementation System	24
Table 8: Types of Big Data Platforms	26
Table 9: Distribution of Data on 16 Platforms	37
Table 10: Centers and Data Products of the 10 Big Data Platforms	49
Table 11: Status of Workers Participating in Data for AI Learning (Unit: number of people)	55
Table 12: Status of Governance by Data Policy Category in Korea	60
Table 13: Public Data Service by User Type	62
Table 14: Verification System for Handling Expenses for Crowd Workers	69
Table 15: The Four Essential Quality Verification Indicators	70

Acronyms and Abbreviations

AI	Artificial Intelligence
APIs	Open Application Programming Interfaces
CKAN	Comprehensive Knowledge Archive Network
COCO	Common Objects in Context
DCAT	Data Catalog Vocabulary
EAP	East Asia Pacific
ESB	Enterprise Service Bus
ETRI	Electronics and Telecommunications Research Institute
GPU	Graphics Processing Unit
IaaS	infrastructure as a Service
ICT	Information and Communication Technology
IoT	Internet of Things
IRB	Institutional Review Board
ITRC	University Information Technology Research Center
KAIST	Korea Advanced Institute of Science and Technology
KCDC	Korea Centers for Disease Control and Prevention
KESTI	Korea Environmental Science and Technology Institute
KISA	Korea Internet & Security Agency
KISDI	Korea Information Society Development Institute
KLID	Korea Local Information Research & Development Institute
KoDi	Korea Digital Development Program
LIDAR	Light Detection and Ranging
LOD	Linked Open Data
ML	Machine Learning
MNO	Mobile Network Operator
MOIS	Ministry of the Interior and Safety
MSIT	Ministry of Science and ICT
MVNO	Mobile Virtual Network Operator
NBIS	National Basic Information System
NCA	National Computerization Agency
NIA	National Information Society Agency
NIPA	National IT Industry Promotion Agency

ODB	Open Data Barometer
OECD	Organisation for Economic Co-operation and Development
OID	Object Identifier
ORS	OID Resolution System
R&D	Research and Development
RDF	Resource Description Framework
SaaS	Software as a Service
SI	Systems Integration
SME	Small and Medium-Sized Enterprise
SOA	Service-Oriented Architecture
SPRI	Korea's Software Policy and Research Institute

* All currency is in U.S. dollars unless otherwise noted.

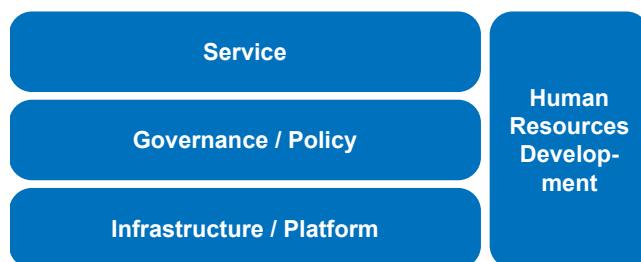
Executive Summary

Over the past few decades, the Republic of Korea has consciously undertaken initiatives to transform its economy into a competitive, data-driven system. The primary objectives of this transition were to stimulate economic growth and job creation, enhance the nation's capacity to withstand adversities such as the aftermath of COVID-19, and position it favorably to capitalize on emerging technologies, particularly artificial intelligence (AI). The Korean government has endeavored to accomplish these objectives through establishing a dependable digital data infrastructure and a comprehensive set of national data policies.

This policy note¹ aims to present a comprehensive synopsis of Korea's extensive efforts to establish a robust digital data infrastructure and utilize data as a key driver for innovation and economic growth. The note additionally addresses the fundamental elements required to realize these benefits of data, including data policies, data governance, and data infrastructure. Furthermore, the note highlights some key results of Korea's data policies, including the expansion of public data opening, the development of big data platforms, and the growth of the AI Hub. It also mentions the characteristics and success factors of Korea's data policy, such as government support and the reorganization of institutional infrastructures. However, it acknowledges that there are still challenges to overcome, such as in data collection and utilization as well as transitioning from a government-led to a market-friendly data policy. The note concludes by providing developing countries and emerging economies with specific insights derived from Korea's forward-thinking policy making that can assist them in harnessing the potential and benefits of data.

Overview of Korea's Data Ecosystem

Figure ES.1: The Four Elements of Korea's Data Policy



Source: Authors

The creation of national fundamental information networks, which began in the late 1980s, established the groundwork for a data ecosystem. The government initiated one of the fastest computer networks in the world before moving its focus from networks to services. Data, an essential resource for the "Fourth Industrial Revolution," grew as a result of these services. With the rise of AI in 2010 onward, the government launched initiatives to actively use this data to complete the incremental shift from

computerization to informatization to smartification. Following that, it advocated for a series of strategies and plans aimed at promoting the national expansion of the data sector and services while also creating an environment conducive to the natural interconnection, exchange, dissemination, and utilization of public and private data. In pursuance of this objective, the government has placed emphasis on infrastructure and platforms, policy and governance, services, and human resource development.

In 2012, the government put the "Big Data Master Plan for Smart Country Realization" into action. The objective was to stimulate the private sector's innovative engagement in the fast-rising data economy and to provide a structured approach to managing public data, which was constantly expanding. Furthermore, it developed the "Big Data Service Activation Strategy" to assist organizations in the development of services that leverage large volumes of confidential data to meet essential criteria (such as fortified institutions, skilled personnel, and a data platform).

¹ Disclaimer: The analysis presented in this policy note is limited to information that is presently accessible and projects that have been finalized as of December 2022, the conclusion of the research examination. After this date, changes to policies and the introduction of new programs are not addressed in this report.

Following years witnessed the deployment of data-driven industrial strategies by the government, with the goal of improving policy implementation. Korea launched the “New Internet Industry Promotion Plan” in 2013 to boost the country’s competitiveness in three critical technologies: data, the Internet of Things (IoT), and cloud computing. The Presidential Committee on the Fourth Industrial Revolution was formed in 2016, with a focus on data, networks (5G), and AI. The government’s “Data Industry Revitalization Strategy,” published two years later, expanded on its intention to develop a comprehensive support system for the whole data life cycle.

Data Policy and Governance

In Korea, data governance falls into several regulatory and institutional layers. The Presidential Special Committee on Data Policy acts as the nation’s data monitoring tower, advising the president on data policy issues and encouraging inter-institutional collaboration on data-related projects and concerns. It reports directly responsible to the president. The Ministry of Science and ICT (MSIT) and the Ministry of Interior and Safety (MOIS) are in charge of establishing data-related policies and promoting new projects, respectively. The National Information Society Agency (NIA) collaborated with both ministries to get these new policies and programs forward.

Unless an exception applies, it is normal to reveal all public data. The “Act on the Promotion of the Provision and Use of Public Data” mandates that each government agency to appoint a Director General of Public Data Provision and the Public Data Strategy Committee to develop public data policy.

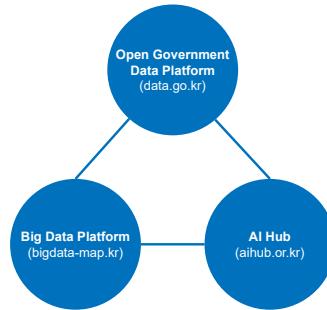
The recently revised “Framework Act on Intelligent Informatization” of 2020 authorizes the development of legislative frameworks that enable standardization and other required provisions for the management, support, and compatibility of data pertinent to private information. Furthermore, the legislation established the NIA as a specialist support organization that campaigns for data dissemination and disclosure for public benefit, as well as the social use of essential private information that meets the criteria for a public good. It also establishes a collaborative structure to ensure the efficient development, gathering, oversight, dissemination, and application of critical data. The passage of the “Data Framework Act” in 2021 creates a market in which rights are guaranteed and data is valued, enabling for its interchange and diffusion. Furthermore, it provides a framework for evaluating the value and quality of data, lays the groundwork for data trading, and aids the training of skilled data traders.

Data Platforms

The Korean government has established three platforms — the Open Government Data Platform, Big Data Platform, and the AI Hub — with the intention of creating a data ecosystem that facilitates the access and organic connection of public and private data.

The onset of the COVID-19 pandemic in 2020 served as a catalyst for the implementation of a “Digital New Deal,” which entailed increased government spending on data collection, processing, and distribution in order to establish a data ecosystem, promote industry development, and generate employment opportunities. The resultant “data dam” amasses information produced by private and public networks, refines and processes it in accordance with established criteria, and subsequently distributes it to businesses, enabling them to generate novel data and develop more inventive services.

Figure ES.2: Korea’s Three Core Data Platforms



Source: Authors

The primary endeavors of this Digital New Deal can be classified into three overarching categories:

- *The Open Government Data Platform.* The aim of this endeavor is to ascertain and divulge information that is readily accessible to the public and possesses substantial value for enterprises. By performing data pre-processing, analysis, and visualization, young individuals have gained experience in data-related fields and obtained gainful employment.
- *The Big Data Platform Project.* The aim is to promote cooperation between the public and private sectors with the purpose of developing big data platforms that collect and distribute industry-specific data in a systematic manner. Furthermore, the objective of this endeavor is to create data centers that will provide these platforms with continuous data. Moreover, it serves as the foundation upon which the data that is to be exchanged as a product of economic value is built.
- *The AI Training Dataset Project.* This undertaking generates datasets that are subsequently made accessible through the AI Hub for use in the training of AI models. Individual collection or creation and annotation of data from sources including text, audio, images, and video requires a significant number of recently unemployed individuals to be provided with minimum wage employment opportunities.

The Open Government Data offers a search function and distributes public data in a variety of formats, including file data, an open application programming interfaces (API), and visualization, to facilitate its use by the public.

Big data platforms enable stakeholders, including consumers and suppliers, to generate new businesses throughout the data life cycle, derive value from data, and exchange data. Big data platforms involve specialized data centers (organizations that systematically generate, construct, and disclose in-demand, high-quality data) that play a crucial role in developing and producing data, managing its quality, and providing support to organizations that actively utilize data.

Additionally, the Korean government intends to develop and make public 1,300 distinct categories of AI training datasets via the AI Hub.

Analyzing Korea's data policies and practices

Several accomplishments have been noted thus far as a result of the mentioned ongoing data policies and the data initiatives implemented under the digital new deal.

Opening Public Data

The content and scope of public data accessibility have been substantially broadened as a result of the strengthening in the Open Government Data platform. Additionally, the revitalization of start-ups and the private utilization of public data have been improved. The public data management system has also been reinforced. Consequently, since the outbreak of COVID-19, diverse private sector services have emerged and Korea's global ranking in terms of public data has increased.

Big Data Framework

A total of sixteen big data platforms have been developed by the public and private sectors to provide the essential data required for policy development and related industries. To adequately support the big data platforms, the data infrastructure was expanded to include 21 centers, led by private companies, in addition to the 108 data centers that were established under the direction of central government agencies, local governments, and public institutions. These centers are presently operational and able to provide the new data in a seamless fashion. These endeavors have contributed to the facilitation of data trading, the establishment of a framework for data dissemination, the support of data analysis, the promotion of standardization and quality enhancement.

The AI Hub

As of this moment, hundreds of millions of training cases have been amassed on the AI Hub, which has experienced a significant surge in visitors and a notable improvement in the efficacy of downloaded data. Additionally, small and medium-sized enterprises (SOEs), startups, students, and individuals with limited access to AI resources developed a developing interest in AI data. The emergence and widespread adoption of AI innovation services commenced with the establishment of AI training datasets and their subsequent accessibility to the private sector. In addition, registration of intellectual property rights and academic research from both the United States and abroad have been initiated, and a substantial workforce is engaged in the development of data.

Characteristics and Success Elements

Characteristics

The data policy and practices in Korea are predominantly risk-taking, social and industrial innovation-focused, and initiated by the government. The role of the Korean government extends beyond that of a regulator to that of a participant in the development, maintenance, and direction of the entire data cycle. The government assumes the social risks necessary for the progression of the data economy by constructing diverse platforms and enacting policies. This proactive stance in Korea's data policy is intended to foster innovation in Korean industry and society while addressing economic and social challenges (such as the COVID-19 crisis).

Success Elements

Several elements have played a role in the accomplishments of Korea's data policy thus far. The prioritization and leadership of the government have facilitated the mobilization of numerous legislative resources and institutional capacities for the implementation of data policies. The reorganization of institutional systems is primarily intended to improve data utilization, resolve systemic barriers to the implementation of the data policy, and address concerns regarding the protection and use of personal information. Additionally, it facilitated the utilization of human and digital resources that had been amassed during the national informatization and digital government transformation processes.

Challenges in Korea's Data Practices

Continued obstacles remain to be overcome. In the present data practices, the following areas are included:

Collection and Generation of Data

In comparison to the foremost data-producing nations, the extent of data accessibility and utilization is still restricted, and substantial gaps persist between platforms. Moreover, significant variation exists in the data utilization capabilities of public sector organizations, as evidenced by issues such as inadequate data quality, discrepancies between data collection and production capabilities, and scarcities in data availability.

Distribution and Trading of Data

Data trading is being constrained by a dearth of distribution channels and demand uncertainties. Furthermore, the source of data is difficult to determine, which results in extremely high data trading costs. Moreover, trade negotiations are subject to limitations as a result of information asymmetry, which encompasses aspects such as price calculations, standardization, and quality.

Analysis and Utilization of Data

Considerable diversity exists regarding the capabilities of data analysis and utilization. Most businesses and small and medium-sized enterprises (SMEs) lack the capacity to utilize data; as a result, the adoption of data throughout society and industries is relatively

sluggish. Furthermore, there is a lack of sufficiency in the foundation for the convergence and combination of valuable data. Finally, consumer requirements are challenging to reflect, and the experience and usability of the data are inadequate.

Future Tasks for Enhancement

The policy note underscored three specific domains that will necessitate enhancement in forthcoming data policy undertakings:

Implementing Efficient Data Governance

In Korea, the governance of data is decentralized. To optimize the efficacy and sustainability of Korea's data policy, the achievement of several critical objectives is of the utmost importance. Establishing an all-encompassing and cohesive data governance framework that integrates the involvement of both the public and private sectors is of the utmost importance in the beginning. Furthermore, it is imperative that a wide range of institutions and stakeholders work together in concert to establish a cohesive vision and objective. In conclusion, across the entire data ecosystem, enforcement and coordination of policies pertaining to data needs to be streamlined and coordinated. For the formation of networks with private businesses, government agencies, and foreign nations, these elements will be crucial. Integrated data management implementation will further necessitate the collaboration of institutions and ministries with an emphasis on data.

Creating Data Platforms with Users and Consumers in Mind

From a supplier-centric standpoint, data disclosure and the development of platforms have occurred. Core and large-capacity data accessibility is restricted. By adhering to the principle that data ought to be in an open state by default, every piece of data would be exposed.

Transitioning from a data policy led by the government to one that is market friendly

In countries that are leaders in data, the growth of the data market is propelled by the data processing market, while the government oversees public data opening, distribution, and trading, as well as virtually every aspect of data policy. In contrast, countries that are leaders in data prioritize large technology companies in their efforts to promote data development and utilization. While data policies driven by the government are stimulating the utilization of data and generating vitality in the ecosystem, it is crucial to assess their long-term sustainability. There is a need for a market-friendly, virtuous cycle paradigm.

The objective is to establish a data ecosystem that is conducive to the market, wherein the private sector can independently generate value from data, while consistently advocating for government assistance and essential regulations to safeguard data. This will necessitate a strategic expansion, an increase in consumer convenience, and improved data utilization. This objective will be further advanced by fortifying the supply system based on the data platform with incentives that motivate both public and private entities to voluntarily contribute and make available high-quality data. Additionally, the data distribution and transaction functionalities will be enhanced to guarantee that users can access, search, and trade with confidence via the data platform. Commencing the process would involve formulating comprehensive strategies—in collaboration with industry, academic, and research experts—that delineate with more precision the domains and responsibilities that warrant government involvement and those that are more appropriately suited for private sector leadership.

An Approach to the Future of Data Policy Making

Through an analysis of Korea's data practices, this policy note offers significant insights and critical factors that would guide the development of effective data policies. These considerations would empower other nations to construct strong data governance frameworks, data ecosystems that are conducive to the market, and the social and technical infrastructures required for data utilization. Some such factors to bear in mind include the subsequent:

- **Timing** - Determining the optimal moment to implement the policy. The announcement of Korea's AI learning data development project coincided with a critical juncture, as it capitalized on the momentum generated by an AI tournament and subsequently intensified efforts considering the COVID-19 pandemic.
- **Agility** - Implementing precise and agile decision making. The swift ratification of the Digital New Deal exemplifies how correct and timely policy formulation by public officials has been one of the most pivotal factors in Korea's ICT development.
- **Engagement** - By capitalizing on the combined intellect of "crowd workers." A recent policy decision by the government permits individuals possessing digital capabilities to engage in data classification activities for the AI training datasets.
- **Quality** - Placing unwavering emphasis on quality. The primary factor that determines the success of the AI learning data endeavor is the caliber of the data that was generated. In recent times, most models necessary for the development of AI-based services have proven inadequate in tackling all challenges. However, by employing the neural network model divulged in the numerous references, these models may prove adequate in addressing significant issues.
- **Privacy** - While sacrificing quality, increase privacy. Data privacy takes precedence over data quality. The "Personal Information Protection Act" of Korea pertains to the stringent oversight of personal data acquisition and utilization.
- **Capability** - Enhancing institutional capabilities in the domains of policy formulation and execution. The significance of an organization like NIA, which has consistently provided backing for government policies and initiatives, was largely inconsequential in the Korean government's ability to execute a nationwide undertaking of this magnitude within such a condensed timeframe.

In summary, this policy note provides advanced analysis, practical recommendations, and insightful perspectives to aid developing and middle-income countries in the establishment of a sufficient digital data infrastructure. Among these are the optimization of AI utilization and the implementation of streamlined data policies. By drawing lessons from Korea's experiences and implementing them in their respective contexts, developing nations could harness the capabilities of AI systems driven by data and promote sustainable development.

I. An Overview of Korea's Data Ecosystem

1.1 Background

The Republic of Korea's data policy was begun in tandem with the development of the country's computer networks. As part of a transition to an information-based economy, the Korean government has not only continuously built computer networks, but, as a national strategy, has also used such networks to identify key online-based services, and then prepared an environment for securing and distributing the databases and data needed to provide such services.

Through the National Basic Information System (NBIS) project, which unfolded in two phases from 1987 to 1996, the Korean government selected and developed five national basic information networks (administration, finance, education, national defense, and public security). The administrative computer network was designed to connect all central administrative agencies and local governments in a single network, and to establish the core systems for governing the nation (such as resident registrations and real estate and automobile transactions), while establishing the necessary administrative database. The Korean government invested \$1.96 billion in the first phase of its development (1987–1991) and \$3.53 billion in the second phase (1992–1996), for a total of \$5.5 billion,¹ an unprecedented level of investment at the time.

The Korean government's aggressive investment in the NBIS moved the market. To operate and maintain these networks, companies specializing in network operation and maintenance, such as KT and Dacom, have emerged. Large corporations such as Samsung, LG, and Daewoo won government-ordered data and service development contracts, and many new systems integration (SI) companies also emerged. These companies accumulated a lot of experience and expertise, in areas ranging from ICT infrastructure to service development, through the NBIS.

Based on the experiences and achievements of the NBIS, the Korean government then made the Broadband Network Project a national priority. The government and the private sector together invested a total of \$16.28 billion in this project between 1995 and 2003. It culminated in the creation of the world's fastest network and its benefits were judged to be approximately 7.3 times greater than the invested amount. More specifically, it led to the creation of a communications equipment and service market valued at \$118.74 billion, and a broadband network market valued at \$35.34 billion; approximately \$7.94 billion worth of value was also created in other areas. There were also ancillary effects, such as the employment of 230,000 people, the production inducement effect of \$20.65 billion, and the production of \$61.95 billion-worth of equipment.² Furthermore, through this project, Korean network operators grew to a level that allowed them to compete with world-class companies in terms of technology, infrastructure, and experience.

Despite having the world's fastest network, in terms of ICT services, Korea was still lagging behind more developed countries. The Korean government therefore shifted its network-oriented ICT policy towards a service-oriented policy and promoted the establishment of online services such as e-government. From 2001 to 2007, the government invested \$952.91

1 Jungsoo Park, Soonae Park, Jung-Ju Lee, and Sun-Ha Kim. (2003). "Review of the Adequacy of the Informatization Budget and Efficient Resources Allocation." University of Seoul Law and Administration Research Institute.

2 Yong-Gwan Jeong & Yoo-Jung Kim. (2004). "An Exploratory Study on Analyzing the Multi-Dimensional Effectiveness of Broadband Network of Korea." National Computerization Agency, Information Systems Review, Vol.6, No.2, December 2004.

million to build 42 e-government services at the national level in two phases. Most of the country's existing e-government services were created during this period.

Such policy efforts resulted in the specialization and diversification of the ICT market, which at that time included network companies that build and operate networks (such as KT, LGU+, and SKT), SI companies that develop services (such as Samsung SDS, LG CNS, and SK C&C), and database companies (such as Altibase, Cubrid, and Tmax). With more companies emerging than ever before, new jobs were created, and competition was intensified, resulting in an overall increase in the nation's ICT competitiveness.

In the process of operating the public services directly developed by the Korean government, as well as the private services that supported them, tremendous amounts of data were accumulated. The government disclosed the data to increase transparency and fully foster the data industry, a key resource in the era of the Fourth Industrial Revolution. In 2010, the Korean government began to pay closer attention to issues related to data as it began to prepare for a new leap forward based on its achievements in informatization. As the informatization paradigm shifted from network and infrastructure centered to software centered, and the importance of data (the basis of software) increased, the Korean government began to focus on actively utilizing data in order to maintain its status as an informatization powerhouse and a global leader in the ICT field.

ICT governance in Korea has undergone as many changes as ICT policies. In the initial (computerization) phase of informatization, the National Infrastructure Network Committee was formed under the direct control of the Office of the President, and the Ministry of Science and ICT³ was established as a ministry dedicated to ICT in 1994. Furthermore, the National Computerization Agency (later the National Information Society Agency) was established to provide professional technical support. With the Korean government promoting various ICT policies as national priorities, the Informatization Strategy Committee and the e-Government Committee were established at the national governance level, and then after 2010, when artificial intelligence (AI) began to be promoted in earnest, the Presidential Committee on the Fourth Industrial Revolution and the Presidential Special Committee on Data Policy were established.

Figure 1: Changes in the Korean Government's ICT Policy

	Infrastructure	Database and Data	Key National Service Strategy	Key governance bodies
Computerization (1980s)	Five major national networks (administration, finance, education, defense, public security)	Basic administrative database (residents, automobiles, real estate, employment, etc.)	Master plan for administrative computerization	National Computing Network Committee, Administrative Computing Network Committee, Ministry of Government Administration, Ministry of Information and Communications (former Ministry of Science and ICT), and NCA (newly established)
Informatization (1990s~early 2000s)	High-speed network (development of high-speed information and communication infrastructure)	National database project *to overcome national financial crisis	Master plan for nation informatization Master plan for e-government	Informatization Strategy Committee, E-government Committee, Korea Communications Commissions, Ministry of the Interior and Safety, and NIA
Smartification (2010~present)	Internet-based new industry (ICBM)	Data dam (public data, big data, AI data) *to overcome COVID-19 national crisis	National artificial intelligence strategy	Committee on the Fourth Industrial Revolution, Special committee on data, Ministry of Science and ICT, Ministry of the Interior and Safety, and NIA

Source: Internal Data and materials of the NIA

Note: * Squares mark the key area of policy by period, and “→” marks the direction of change in the key area.

³ It was established as the Ministry of Information and Communication in 1994 and then changed to the Ministry of Science, ICT and Future Planning (2013-2017) and now became the Ministry of Science and ICT (2017-Present)

In 2012, the National Informatization Strategy Committee⁴ announced a policy called “Big Data Master Plan for Smart Country Realization,” intended to systematically manage increasing national-level data and support the creative use of the private sector to boost Korea’s future competitiveness. It also established and announced the “Big Data Service Activation Strategy” to meet the demands of companies developing services using large-scale private data in areas such as telecommunications, finance, and the internet. These plans support the development of a data platform, skilled workforce development, and data-related institutional improvements.

Table 1: Overview of the Big Data Master Plan and Big Data Service Activation

Classification	Key Contents
Development of foundation for use	Development of common facilities for sharing and utilizing big data within the government, and establishment of a big data support center.
Technology development	Preparation of technology development roadmap, support for key technology development, and strengthening of platform competitiveness.
Legislative maintenance	Preparation of personal information protection measures and system improvement.
Training and capacity development	Industry-academia joint research and development.

After announcing the above two strategies and presenting future policy directions related to Korea’s data, the government established and announced data-based industrial strategies to enhance the execution power of policies centering around the Ministry of Science and ICT. First, in 2013, the government devised the “New Internet Industry Promotion Plan” to foster the three largest technologies—data, internet of things (IoT), and cloud. These were considered new growth engines of ICT, and interest in the new internet industry was rising. However, Korea was not a leader in these sectors, so the government devised its “Internet New Industry Fostering Plan,” to rapidly raise the country’s competitiveness in these technologies to a level comparable to that of the leading countries to spur the development of new businesses.

Table 2: Overview of Plan to Foster New Internet Industry

Classification	Key Contents
Formation of foundation	<ul style="list-style-type: none"> Preparation of laws and systems to promote the spread of new internet industries and secure user reliability. Promotion of R&D considering the need to secure and commercialize source technologies. Fostering insightful specialists by combining humanities and new internet industry technology.
Creation of market	<ul style="list-style-type: none"> Creation of a sustainable ecosystem by expanding IoT services, strengthening the data and ICT resource utilization system, and supporting commercialization efforts by companies. Promotion of leading projects to activate local industries, enhance public safety and convenience, and strengthen IT-related competitiveness.
Strengthening of competitiveness	<ul style="list-style-type: none"> Development of corporate support infrastructure such as open data analysis utilization center and global future internet demonstration environment. Creation of a foundation for growth of small and medium-sized enterprises (SMEs) by providing a new service development environment, commercialization consulting, and support for securing intellectual property rights.

Furthermore, with the advent of AlphaGo in 2016, the government formed the Presidential Committee on the Fourth Industrial Revolution, to supervise and coordinate relevant tasks in preparation for the advent of the era of AI. It designated data, network (5G), and AI as its

⁴ As the presidential committee control tower for the supervision and coordination of the national informatization projects, its role includes providing strategic support for key national tasks through informatization and leading the knowledge information society by discovering a future-oriented informatization agenda.

priority areas, seeking to focus on national competencies. This was not simply a transfer of responsibilities over data from the ministries in charge of ICT to the Presidential Committee, it also showed the will and determination of the government to treat data as a higher priority and push a policy direction that placed a greater significance on data. The government's determination was reflected in its "Data Industry Revitalization Strategy," announced in 2018. At the core of this strategy was a blueprint for preparing a comprehensive support system throughout the data life cycle (collection and construction, storage and distribution, analysis and utilization).

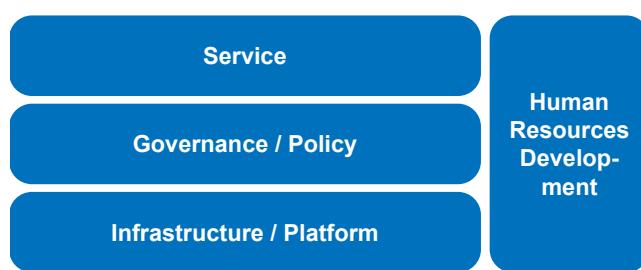
Table 3: Overview of the Data Industry Revitalization Strategy

Classification	Key Contents
Paradigm shift of the data usage system	My Data, Data Safety Zone.
Innovation of entire cycle for data value chain	Development of data for AI learning, fostering big data specialized centers, etc.
Laying a foundation for fostering the data industry	Securing leading technology for big data that is at 90% of the level of developed countries or higher by 2022 and fostering a professional work force.

As explained above, the government promoted a policy to facilitate the data industry and services at the national level, as well as a policy to fully disclose public data. Since the 1990s, the government has consistently made investments in this area, with the catchphrase "We were behind for industrialization, but will get ahead for informatization." As a result, since the mid-2000s, Korea has maintained the top position in the e-government category, as assessed by various international organizations, including the United Nations and the Organisation for Economic Co-operation and Development (OECD). The government was able to gather vast amounts of administrative data through various e-government systems built with huge investments. Going beyond simply accumulating this data, the Korean government publicly discloses the data to increase administrative transparency and refines the data to ensure that companies can use it for business purposes, contributing to the activation of the nation's data economy.

1.2 Key Details of Korea's Data Policy

Figure 2: The Four Elements of Korea's Data Policy



Korea's data policy is largely divided into four elements: infrastructure and platform, policy and governance, service, and human resources development. In this section, the main policies developed and promoted by the government are explained in greater detail, along the lines of these four elements, in order to show how Korea's data policy has changed, and where it is focused.

Source: Authors

1.2.1 Development of Data Infrastructure and Platforms

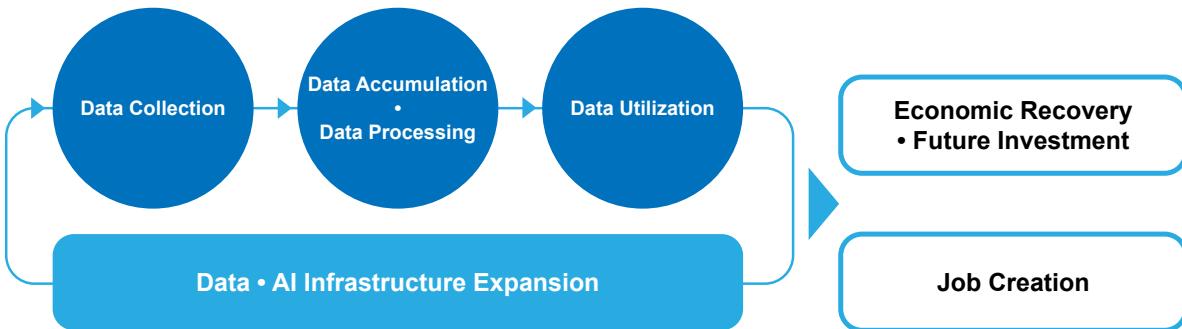
The core of the Korean government's data policy was to create an environment in which public and private data could be organically linked, circulated, traded, and finally utilized. In this process, the government built data that was urgently needed for companies and ultimately tried to establish the data ecosystem in which public and private data could flow freely.

The government has developed a triangular data platform system, comprised of three platforms (Open Government Data, Big Data, and the AI Hub) that together form the basis of a data ecosystem in which public and private data are organically linked and flow without blockage.

Working from the belief that creating a virtuous cycle environment for the overall data ecosystem will determine the success or failure of the data economy, the government endeavored to provide support across the entire data ecosystem through the three largest data platforms. It further prepared for the data economy era by disclosing existing data and building a platform that can collect and build essential data in the era of artificial intelligence, or trade and distribute private data. In particular, the government has used the national crisis caused by the COVID-19 pandemic as an opportunity to lay a foundation for the data economy at the national level.

In 2020, the COVID-19 pandemic set back the Korean government's economic growth plans. But rather than reducing investments in data, the government increased investments designed to realize a data economy early in order to overcome the economic crisis caused by COVID-19 and secure the nation's future growth driver by pursuing a "Digital New Deal" centered on the development of a "data dam." That is, the government chose to invest heavily in the collection, processing, and distribution of data to build the foundation for a data ecosystem and lay the groundwork for future growth industries, while also actively investing in the data dam to secure the work force for data and create jobs. The dam collects the data generated through public and private networks, refines and processes it according to standards, and provides it to companies, allowing them to develop more innovative services and create new data.

Figure 4: Korea's Digital New Deal: Data Dam

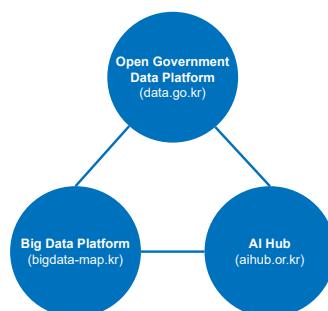


Source: Authors

The core projects of the Digital New Deal can be broadly divided into the following three categories:

- *The AI Training Dataset Project.* This builds on datasets used to train models when developing artificial intelligence and discloses them through a platform called "AI Hub." To create a training dataset, data from sources such as images, video, text, and voice must be collected or created one at a time and annotated to ensure that AI can recognize each object. The annotation requires a large-scale workforce, and thus was able to provide minimum-wage work opportunities for many newly unemployed people.
- *The Big Data Platform Project.* This is designed to engage the public and private sectors to collaboratively build the big data platforms that collect and disclose the data necessary for each industry and to build the centers that continuously provide data to these platforms. It

Figure 3: Korea's Three Core Data Platforms



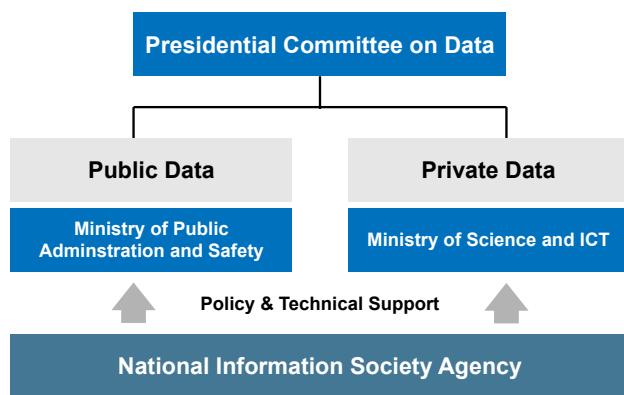
Source: Authors

addresses the problem of data shortage by providing the grounds for distribution of private data with public value and promoting the use of data throughout the industry. Notably, this project goes beyond simply collecting, processing, and disclosing public and private data, and provides (as a platform that can distribute and trade data between participating institutions) the foundation for the data to be traded as a product with economic value.

- *The Open Government Data Platform.* Its purpose is to quantitatively expand the disclosure of public data held by the central government, local governments, and public institutions, by discovering and disclosing the high value data that companies can use. The government introduced a youth internship system to accomplish this task, primarily employing young people in their 20s and dispatching them to public institutions. Their work involves pre-processing data and analyzing and visualizing it at the institution's request. This enabled public institutions to disclose data previously difficult to disclose and gave young people (who had difficulty finding other work) secure jobs as well as data-related knowledge and experience.

1.2.2 Data Governance: Organizational Structure and Legal System

Figure 5: Korean Government Data Organizational Structure



Source: Authors

Korea's data governance has a three-tiered structure. At the top is the Presidential Special Committee on Data Policy, operating directly under the president. As the nation's data control tower, the committee determines the directions of the nation's data policies and plays a role in coordinating data-related projects and issues between institutions.

Next are the Ministry of the Interior and Safety (MOIS) and the Ministry of Science and ICT (MSIT), which develop the data-related policies, request decisions from the Presidential Special Committee on Data Policy, and create and promote new projects. MOIS oversees policies and projects related to public data produced

and held by the central government, local governments, and public institutions. MSIT oversees policies and businesses related to private data produced by companies.

Last, there is the NIA, which supports MOIS and MSIT in developing data-related policies and promoting businesses. By supporting both organizations, the NIA bridges the gap between public and private data policies and prevents any possible conflicts between the two areas.

a. Public Data

According to the "Act on Promotion of the Provision and Use of Public Data," the public data to be disclosed includes all original data generated and owned by the central government, local governments, and public institutions. However, heads of public institutions are allowed to ask MOIS to exclude specific data from the disclosure list if it would compromise public and national safety, such as personal and national security information. To prevent the abuse of exceptions to avoid data disclosure, MOIS reviews exclusion requests submitted by heads of institutions and decides whether to grant them. The data disclosure obligation was strongly underscored even for excluded data; however, procedures were developed for the public to separately apply for its disclosure.

The public data act also compels MOIS to devise a "Basic Plan for the Activation of Public Data Use" to promote the efficient provision and use of public data every three years. Accordingly, the heads of the ministries of the central and local governments must establish

an implementation plan regarding the content, method, and timing of data released each year. MOIS evaluates how well these plans are meeting their objectives on an annual basis.

Table 4: Key Contents of the Act on Promotion of the Provision and Use of Public Data⁵

Classification		Contents
General Provisions	Purpose and Definition	<ul style="list-style-type: none"> Guarantees the public's right to use public data. Provides permission of delivery or access in machine readable form (a state in which the contents of public data can be processed [i.e., revised, converted and extracted] by software).
	Basic Principles	<ul style="list-style-type: none"> Public institutions must take necessary measures to support the universal expansion of the right to use public data and ensure the commercial use of public data.
Development of Policy	Public Data Strategy Committee	<ul style="list-style-type: none"> Formation of committees under the Prime Minister's Office to deliberate on major policies, such as the development of and alterations made to the basic plan and its scope.
	Implementation of Basic Plan and Evaluation of Operational Status	<ul style="list-style-type: none"> Development of the basic plan for public data provision and use (three-year cycle; MOIS) and development of the implementation plan (yearly; heads of central and local governments). Evaluation of the operational status of public institutions' data provision (MOIS).
	Director General of Provision and Utilization Support Center	<ul style="list-style-type: none"> Appointment of a person to oversee the work of institutions at each level. Establishment of a center for work support.
Formation of Foundation	Public Data Subject to Provision, etc.	<ul style="list-style-type: none"> Subject to provision: Data under jurisdiction (confidential and copyright-protected information excluded). Deliberation (Strategy Committee): Deliberation and resolution on the list of data subject to provision. Data registration: Public data portal (public institution). Establishment of institutions of provision: Maintenance of data in machine-readable form and procurement of various types of provision methods, etc. (public institutions). Infrastructure support: Portal operation, quality control, standardization, and data maintenance support (MOIS).
Provision of Data	Data Provision and Public Data Dispute Mediation Committee	<ul style="list-style-type: none"> Provision method: in person or through portal. In the event of a refusal or interruption of data provision, a committee must be formed for dispute mediation.
Supplementary Rules	Exemption	<ul style="list-style-type: none"> Exemption of public officials from civil and criminal liabilities for damages to users due to interruption of data provision, etc.

The “Act on Promotion of the Provision and Use of Public Data” provides for the pan-governmental Public Data Strategy Committee to devise public data policies and evaluate the progress made toward meeting the goals. The prime minister and a private-sector expert are the committee’s co-chairpersons; the majority of the members represent the private sector.

To guarantee the policy’s implementation, the head of each public institution appoints a director general (hereinafter the Director General of Provision for Public Data) and a person in charge of affairs related to the provision and use of that institution’s public data (Article 12 Paragraph 1 of the Act). The Public Data Utilization Support Center was also formed under the NIA, to support the promotion of data policies.

⁵ Seok-jin Eom. (2022). 2022 KIPA Module Series. “Legal and Institutional Arrangements for Digital Government”. (https://www.kipa.re.kr/synap/skin/doc.html?fn=FILE_000000000162450&rs=/convert/result/201512/)

Table 5: Public Data Act of Korea: Provision and use of public data⁶

Classification	Contents	Other
Public Data Strategy Committee	Deliberates on and coordinates major government policies and plans related to public data, and checks and evaluates progress on implementation.	
Ministry of Public Administration and Security	Supervises the overall business related to the management, sharing, disclosure, utilization, etc., of public data.	Policy
Public Data Utilization Support Center	Supports the efficient provision of public data and activation of use.	
Public Data Provision Dispute Mediation Committee	Mediates disputes regarding the refusal and suspension of provision of public data by public institutions.	Support
Director General of Provision for Public Data	Oversees an institution's public data provision and promotion of use.	Execution

b. Private Data (Big Data Platforms and AI Hub)

The Big Data Platform Project and the AI Training Dataset Project started in 2017 and 2019, respectively. The Korean government overhauled and revised the National Informatization Framework Act to the Framework Act on Intelligent Informatization in 2020 and enacted the Data Framework Act in 2021 to prepare some grounds for data-related policies. Through this act, the data-related policies that serve as the basis for intelligent informatization were prepared, along with a legal foundation that enables the government to provide intensive support.

Establishing a virtuous cycle ecosystem of data production, collection, distribution, and utilization requires the comprehensive integration of public and private data. The legal grounds were established for the standardization and other provisions needed to manage, support, and secure compatibility between data. In addition, the act promotes data disclosure and distribution for the public good as well as the social use of important private data with the characteristics of a public good; establishes a cooperative system for the efficient production, collection, management, distribution, and utilization of necessary data; and designates the NIA as a specialized support organization (integrated support center).

The Data Framework Act was enacted to promote the distribution and utilization of data to realize the data economy. It provides that data may be traded and distributed by acknowledging the value of data and creating a market in which rights are guaranteed. Furthermore, the act provides a foundation for measuring the quality and value of data, prepares a basis for data trading, and supports the training of professional data traders.

Under the act, various mechanisms induce the broad participation of the private sector. The Data Policy Committee (which oversees national data policy) and the Dispute Mediation Committee (which mediates disputes in the process of data production, transaction, and utilization) were formed, and authority and roles were granted to various organizations to promote data trading and utilization (e.g., a data exchange was established, and the Dispute Mediation Committee was granted the authority to properly assess the value of data and mediate disputes arising from the unauthorized acquisition, use, or disclosure of data, or in the process of data production, trading, and utilization).

⁶ Sungsoo Hwang and Joon Mo Abn, (2022), "Digital Government and Public Data Act of Korea"

Table 6: Key Contents of the Data Framework Act

Classification	Key Contents
General Provisions	The purpose of the act is to create economic value from data and lay the foundation for the development of the data industry, thereby contributing to improvements in people's lives and national economic development; and to define related terms such as data.
Development of basic plan	The government devises the basic plan for the promotion of the data industry every three years to promote data production, trade and utilization, and lay the foundation for the data industry.
Protection of data assets	Protect data with economic value ("data assets") created by significant human and material investment and effort. *Unauthorized acquisition, use, disclosure, provision to other parties, removal of technical protection measures applied to data assets without legitimate authority, etc., are prohibited.
Data valuation support and quality control	Prepare data valuation techniques, valuation systems, quality certification targets and standards, and implement the designation of a valuation agency and a quality certification agency in charge of related tasks.
Data trader	A person with expert knowledge in data trading can register as a data trader with MSIT, which provides such traders with the information and training necessary to perform data trading tasks.
Data industry promotion and data business report	Data trading business operators, data analysis providers, etc., must report to MSIT; MSIT and related central administrative agencies can provide them with the necessary financial, technical support, etc.
Data industry support	In the event of the activation of data-based industry, strengthening of data-related capabilities of companies, support for commercialization, etc., SMEs shall be considered a priority when implementing various data support policies, and partial support for necessary costs such as data trading and processing will be provided.
Fostering of professional work force	MSIT and MOIS prepare policies for fostering data experts, and MSIT designates and supports professional human resources training institutions.

Table 7: National Data Policy's Implementation System

Organization	Role(s)
National Data Policy Committee	<ul style="list-style-type: none"> Oversees the public and private data policies (chaired by the prime minister). Devises the basic plan and reviews the overall and coordination plans related to data industry promotion. Improves policies and systems related to data production, trading and utilization.
MSIT	<ul style="list-style-type: none"> Performs overall planning, coordination and management of national data policies.
Formation of the Data Dispute Mediation Committee	<ul style="list-style-type: none"> Mediates disputes concerning the production, trade, and use of data.
Executive institutions	<ul style="list-style-type: none"> Integrated Support Center (NIA) Professional support organization Data exchange Agency specializing in combining pseudonymous information Data Safe Zone Data valuation agency

1.2.3 Data Services

a. Open Government Data Platform

The Open Government Data Platform is an integrated platform that holds the public data created or acquired and managed by public institutions. It provides the data in various ways to ensure that people can conveniently use it, such as via file data, open API, and visualization, and offers a search function.

The NIA's Public Data Utilization Support Center oversees planning, developing, and managing the platform. It expands the disclosure of high-quality public data at the point of contact between the public and the private sectors and comprehensively supports its creative use by the private sector.

Key roles of the public data utilization support center are:

- Survey and study on policies and systems related to the provision and use of public data
- Survey and analysis of statistics related to the provision and use of public data
- Management support, including the process of public data provision
- Promotion of the use of public data and start-up support
- Support for private sector and international cooperation in public data, and support for a fact-finding study on the development and provision of duplicated and similar services
- Support for securing legitimate permission to provide and use public data, such as copyright support
- Support for the registration of a public data list and management of registered information
- Public data list support and list information service subject to provision
- Promotion of the establishment, management, and utilization of the Open Government Data Platform
- Support for quality assessment, evaluation, and improvement of public data
- Support for the standardization of public data
- Support for the reorganization of the public data provision form and establishment of a provision method
- Education and training related to public data
- Consultation on public data provision or use and act on behalf of data provision
- Operational support, etc., for the Dispute Mediation Committee

b. Big Data Platforms

A big data platform is a process environment intended to support a series of processes for extracting value from data (from data collection all the way through to storage, processing, analysis and visualization). To this end, the big data platform must be equipped with scalable, large-capacity processing capabilities; heterogeneous data collection and integrated processing functions; prompt data access and processing functions; mass data storage management functions; and massive heterogeneous data analysis functions. This allows stakeholders such as suppliers and consumers to create new businesses across the overall data process, and to trade data.

Table 8: Types of Big Data Platforms

Type	Details
Collection platform	A platform that collects data from sources such as blogs, bulletin boards, news, etc., and processes and standardizes it to ensure that it can be used.
Marketplace platform	An open-format platform that supports sharing, distribution, and transaction between consumers and suppliers by selectively providing functions such as data collection, processing, verification, payment, and consulting.
Analytical platform	A platform that provides data analysis and visualization tools and an analysis environment to ensure that user data and data provided by the platform can be utilized for analysis.
Data science platform	A platform that provides tagged data sets that can be used for an AI environment and AI learning, such as NLP and AI learning algorithms.

Specialized data centers (entities that systematically produce, build, and disclose high-quality data that is in high demand) participate in the big data platforms. They play an important role in producing and developing data, while managing its quality and providing it according to a set period. Furthermore, they support the companies that actively utilize data to ensure it can be utilized in the market and that a big data ecosystem may be created.

c. AI Hub

Recognizing that the amount of data used to train AI models is directly related to performance, securing high-quality and large-scale AI training datasets for each field is essential for the spread and development of AI technologies. However, the large amount of time and cost required for data collection and processing is a barrier to the introduction and diffusion of artificial intelligence in SMEs, startups, and universities. Accordingly, the demand for quantitative and qualitative expansion of AI training datasets suitable for domestic conditions has rapidly increased in public and private research and in the development of artificial intelligence.

The Korean government plans to build 1,300 types of AI training datasets by investing \$1.98 billion by 2025 through the data dam and to disclose them through the AI Hub. It has prepared a mid- to long-term roadmap for building data in fields that are being strategically nurtured (such as Korean language, video, and images; healthcare; transportation; logistics; disaster; safety; environment; and livestock and fisheries) and is promoting business.

Figure 6: Mid- and Long-Term Plan for AI Training Data Project

Infrastructure Technologies	Create universally usable AI training data such as Korean, videos, and images that can be used in various areas and situations
Strategic Area 1	Create AI training data that can promote public interests in the areas of healthcare, transportation, and safety, and spread AI technologies fast
Strategic Area 2	Create AI training data in the areas that can accelerate innovation in industries

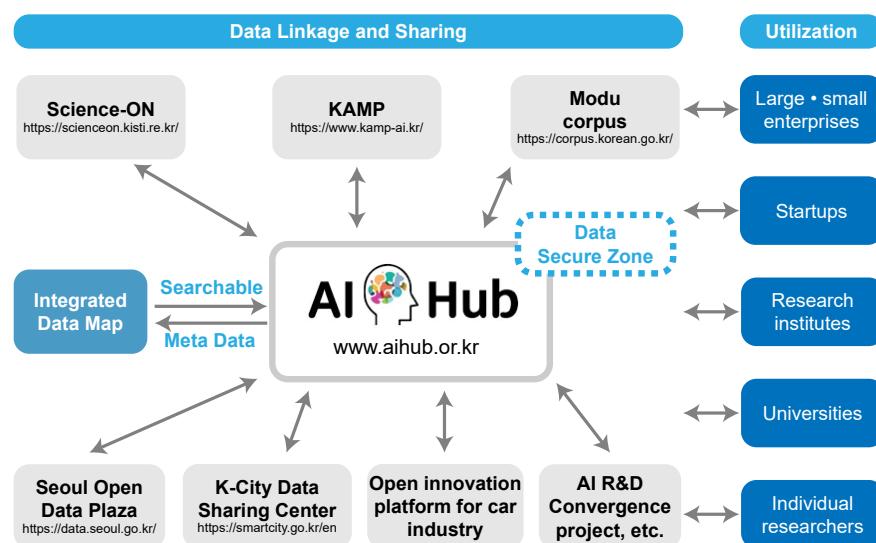
Source: Authors

The data developed through this project is being provided through the AI Hub, and as of April 2022, 191 types of data have been disclosed and provided. Furthermore, the data derived from developing services using the disclosed data have also been accumulated and disclosed again via the AI Hub, thereby creating an environment in which the data can be continuously updated and virtuously circulated.

Moving forward, the plan is for the AI Hub to allow data to be posted on a company's platform, to ensure that it can be modified, reprocessed, and spread, and that results secured through the

data will also be shared. In addition to data collection, accumulation, and provision functions, it is planned for the AI Hub to hold a regular competition (leader board) focused on core data; publish education and practice-related materials; and strengthen its community functions, creating an environment in which everyone can easily utilize data.

Figure 7: Conceptual Diagram for the AI Hub Service



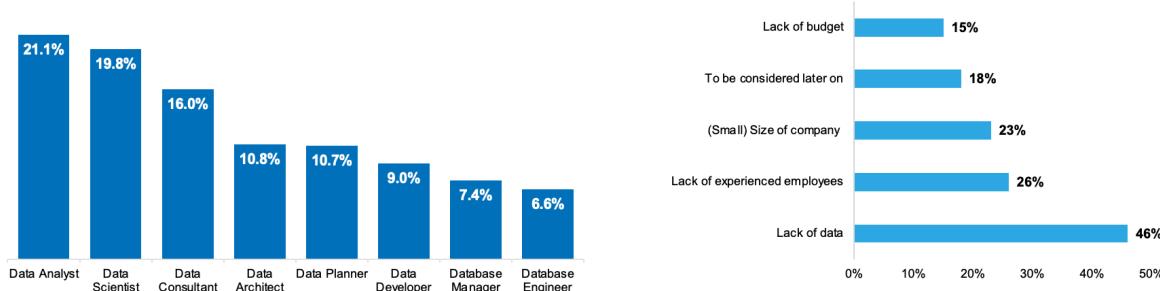
Source: Authors

1.3 Training and Fostering of Skilled Professionals

In 2018, Korea had subpar data technology and industrial infrastructures, including a shortage of specialists. The government's data personnel training policy is therefore a major component of the "Data Industry Revitalization Strategy" announced that year.

According to Korea's Software Policy and Research Institute (SPRI), from 2018 to 2022, 17,073 people were sought for big data positions across the country, but only 14,288 people were expected to be available to fill those positions, representing a deficit of 2,785 people.

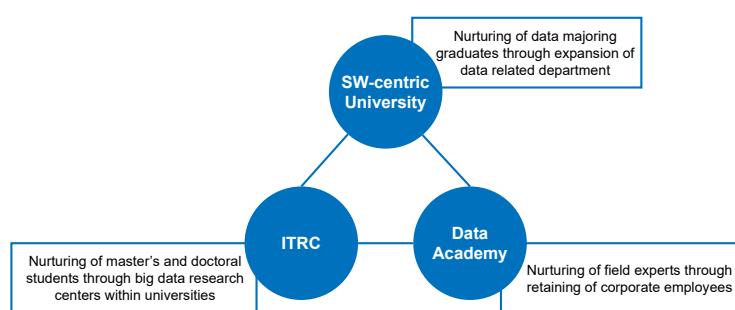
Figure 8: Rate of Big Data-related Labor Shortage and Difficulties of Big Data-related Companies⁷



Source: Ministry of Science and ICT. (2019)

To address the issue of a shortage of data experts, the government proposed three major directions, along with a plan to foster 50,000 data experts consisting of advanced young

⁷ Ministry of Science and ICT. (2019). "The 2018 Status Survey on Data Industry."

Figure 9: Korea's Data Workforce Development System

Source: Authors

government has also provided financial support to universities with strengths in software that have been selected to help develop competitive practical talents, including the development of non-software majors into convergence talents possessing both subject-area knowledge and software skills through software education tailored to demand.

A total of 36 universities (6 in 2017, 10 in 2018, 10 in 2019, and 9 in 2021) were designated as software-focused universities and provided a total of \$258.16 million through 2021. The number of software major-related courses was increased to 4,053 in 2020 from 503 courses in 2015, and the number of software specialists also significantly increased, from 889 people at eight universities in 2015 to 4,918 people in 2020, satisfying the demand for software specialists at the national level. The government's policy has provided diverse incentives for universities to support talented students majoring in data analysis.

Second, the University Information Technology Research Center (ITRC) Fostering Support Project (which the government has backed since 2000 to foster skilled professionals in advanced ICT research) was to expand the number of big-data specialized research centers from one in 2017 to six by 2022, to nurture data scientists. The government has supported university research and development (R&D) on core technologies that will lead the national economy and foster high-quality professionals in ICT by designating ITRCs at universities for each field of technology, establishing an R&D environment, and funding projects to help support research costs for master's and PhD students.

Through this project, 15,841 students received ICT-related master's and doctoral degrees from 2000 to 2020, and 12,602 SCI-level thesis papers, 5,359 patent registrations, and technology fees of \$38.1 million were generated. Researchers from 60 ITRCs installed in 32 universities across the country are now actively participating in cutting-edge research projects.

Third, the Big Data Academy was established in 2013 to help students grow into experts in the data field by providing training to industry workers with practical experience based on corporate demand.

The academy offers three programs to train experts in big-data planning, technology, and analytics. Furthermore, in response to the emergence of new technologies and new demands from companies, three new courses (data visualization expert, data trading brokerage expert, and data processing expert) have been added. By 2020, 86 training courses were offered—producing 2,073 practical data experts since the launch, and 398 pilot projects carried out to help companies address issues in the field.⁸

data experts and practical data experts in its "Data Industry Revitalization Strategy."

First, it sought to support data analysis programs at software-focused colleges and universities. Since 2015, the government has operated a software-focused university with the aim of innovating tertiary education to foster software specialists and enhance software competitiveness. The

⁸ Korea Data Agency. (2022). "Big Data Academy." (<https://dataonair.or.kr/bigdata>)

1.4 A Multi-Platform Data Ecosystem

1.4.1 Open Government Data Platform

a. Production/Collection and Processing/Operation

The government selects data to be made available to the Open Government Data Platform from among the data directly produced by individual public institutions or data purchased externally, including copyrights. Data disclosed to the platform are produced in a machine-readable form, allowing some data contents or structures to be checked and otherwise processed (e.g., corrected or extracted) through the platform's services. The machine-readable data stages are divided into five phases and the goal is to disclose data with the last fifth phase.

Figure 10: Machine-Readable Data Stages

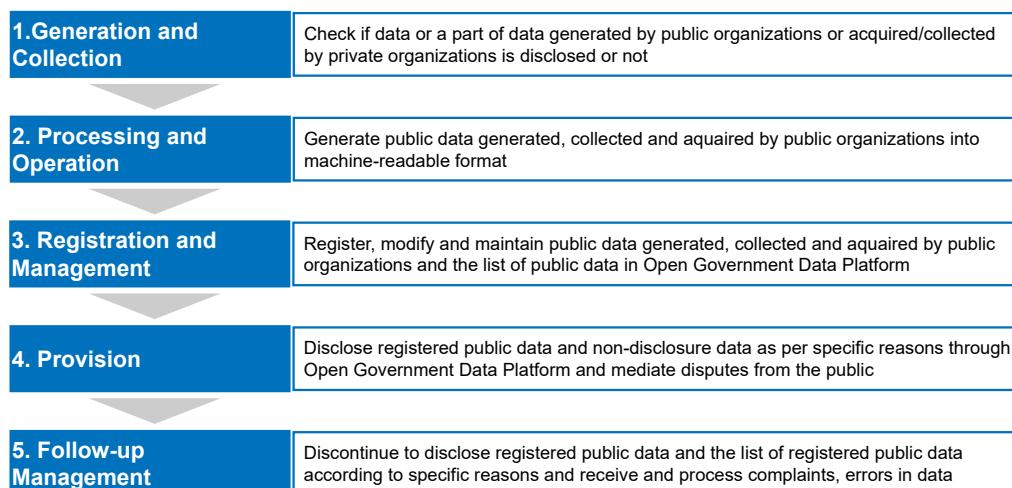
Classification	Phase 1	Phase 2	Phase 3	Phase 4	Phase 5
Forms for machine-readable	Unstructured format	Minimum satisfactory format		Open format	
Characteristics	Data can be read only with specific software	Data can be read, edited, and converted with specific software	Data can be read, edited, and converted with all software	Data structure describing data attributes and relationship based on URI	Data can be connected and shared through the Internet
Examples	PDF	HWP, XLS, JPG, PNG, WMV, MPEG, MP3, SWF	CSV, JSON, XML	RDF	LOD

Source: Authors

b. Data Registration and Provision/Disclosure

Data subject to disclosure are registered with the Open Government Data Platform and disclosed to the public, via direct data download, open API, or linked open data (LOD). The open API method is useful for large capacity dynamic data that are frequently updated; the LOD method is suitable for cases characterized by a lot of data that are unlikely to change, such as names.

Figure 11: Disclosure Stages of Public Data



Source: Authors

In disclosing data through the Open Government Data Platform, data collection, creation, processing, and operation are performed through the institution's internal system area, and data registration management, provision, and follow-up management are conducted through the institution's external system area. Because they are processed and disclosed to the private sector, the security compliance requirements for each are provided separately.

c. Support for the Utilization of Disclosed Public Data

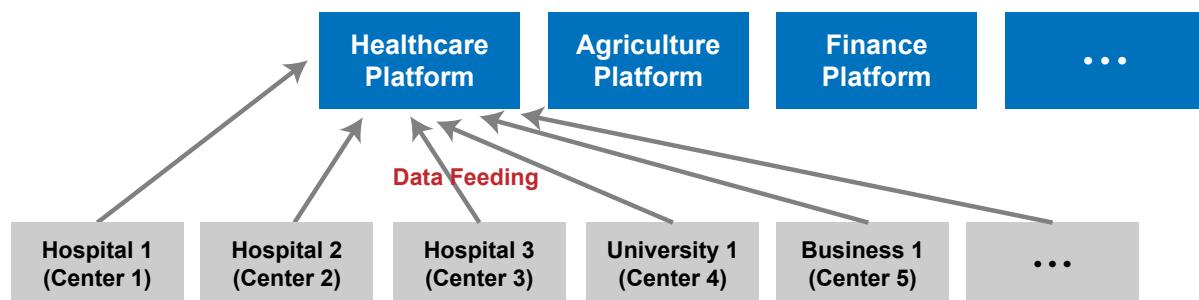
To support the private use of public data disclosed through the Open Government Data Platform, MOIS developed the startup support project using of public data, in cooperation with each ministry. MOIS assists startups to develop new business models by utilizing public data; supplies start-ups' needs for elements that enable commercialization, such as space, funds, infrastructure, and consulting; provides services such as consulting on sales expansion; and offers one-stop assistance with investment attraction and overseas expansion. Furthermore, the Pan-Ministry Public Big Data Utilization Startup Contest is run to discover and promote start-up ideas and business models using public data.

1.4.2 Big Data Platforms

The big data platforms were developed with a focus on distribution and trading, collecting important national data (including private data) by field, storing them systematically, and enabling them to be traded at a fair price. However, they have evolved from simply selling and purchasing data to offering analysis and utilization of important data directly on the platforms.

Accordingly, the platforms provide a forum for new insights and the creation of new business value by engaging various stakeholders such as suppliers, consumers, and intermediaries in the overall process of data collection, storage, processing, and management, as well as data trading. The platforms provide an open, friendly, and sharing environment that enables the public and private sectors to collaboratively collect, produce, and build the high value data that has been neglected in the past.

Figure 12: Conceptual Diagram of the Roles of Big Data Platforms and Centers



Source: Authors

a. Data Collection and Storage

The government selects and supports data providers that facilitate the collection, analysis, distribution, and utilization of important national data, including private data, by sector (e.g., healthcare, agriculture, and finance). Data providers receiving financial support from the government must perform functions such as data quality and security management; collection and linkage system-related establishment and operation; data alliance⁹ participation; and working council composition and operation, as well as distribution portal-related operation and advancement.

⁹ Governance of public-private cooperation to promote the utilization of big data and create an innovative ecosystem.

b. Data Trading (Distribution)

The data center operator that trades the collected data should systematically produce and build the high-quality data that are in high demand and distribute them through the platform. Data center operators perform functions such as data production and construction; data quality management and compliance with the provision cycle; activation of data utilization companies' creation of big data utilization ecosystems; and participation in governance.

c. Data Analysis and Utilization

The big data platforms support an environment that can process and analyze large-scale data to ensure the data can be easily utilized by everyone—individuals, businesses (startups, small businesses, etc.), universities, and research institutes, as well as medium-sized and large companies.

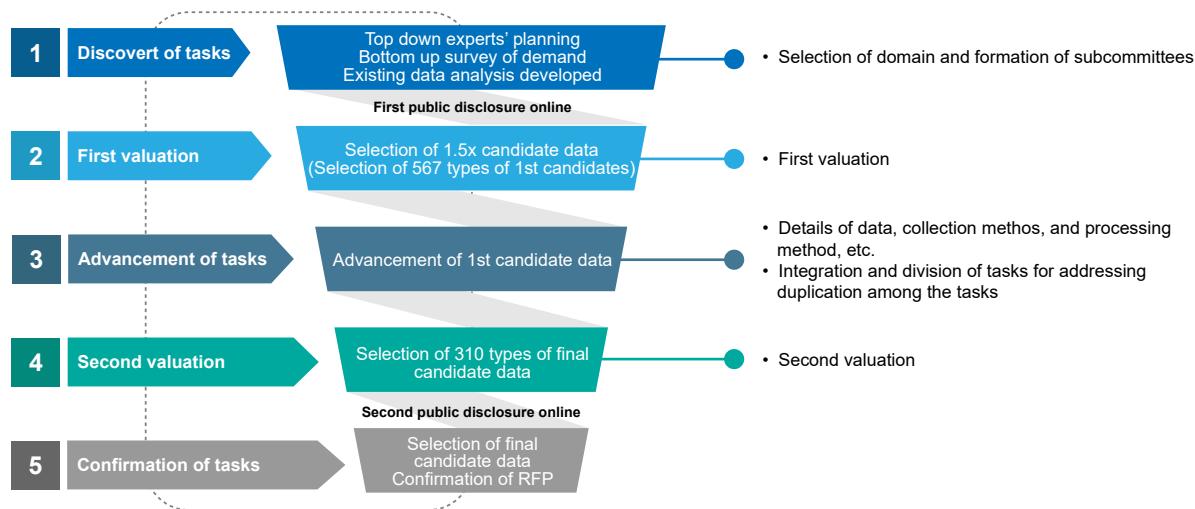
1.4.3 AI Training Datasets

In 2017, when the AI Training Dataset Project was launched, the goal was to quickly build the AI training datasets and provide them to the private sector so that SMEs (which didn't have the resources to produce datasets themselves) could use them to develop AI technologies and services. The government has now invested \$1.59 billion in the AI training datasets, and has expanded the project beyond simply supporting companies, to serve as a foundational business that will be a stepping stone for future growth by creating a national AI foundation.

a. Data Collection and Storage

The decision on which AI training datasets will be created is made in two ways: the AI Training Dataset Planning Committee, which consists of AI experts and domain experts from academia, industry, and R&D institutions, and develops state-of-the-art datasets; while other datasets are created based on the result of surveys by companies and business-related associations.

Figure 13: Data Development Process for AI Learning



Source: Authors

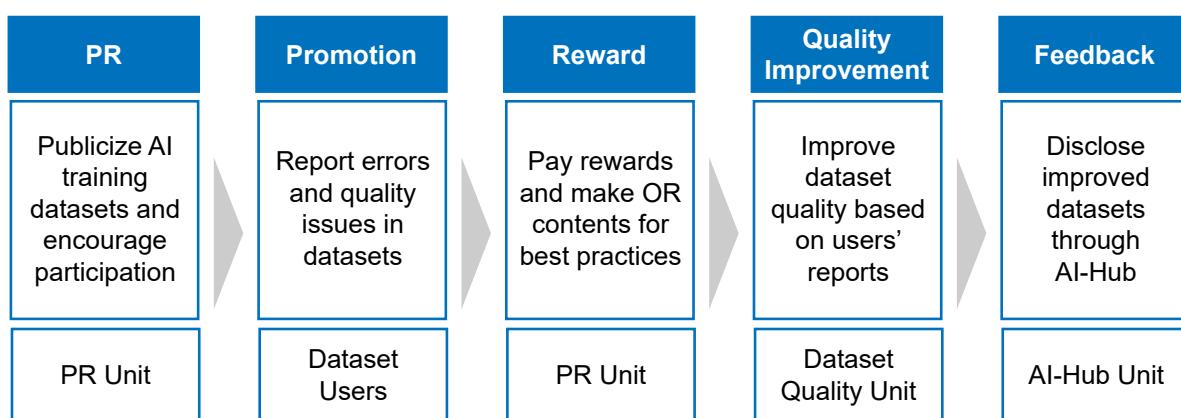
Once candidate data are collected from the committee and surveys, the final task is the selection of the datasets to create through the valuation and in-depth evaluation by the committee. After that, institutions are selected through the open competitive bidding process to create the final selected datasets. The institutions can create the AI training datasets by using high-quality data that they already have or from scratch. At this point, data capable of developing AI services are collected by simultaneously removing bias and securing diversity and by

evaluating whether the quality level required for AI development is satisfied by consulting the AI developers. The collected data are processed as AI training datasets by crowdsourced data labelers.

b. Data Disclosure and Opening (Distribution)

The data collected and processed are subject to quality verification according to diversity, syntax and semantic accuracy, and validity standards; it is also determined whether the data are publicly available through de-identification of personal information, etc. Finally, the data quality and validity are verified by an AI-specialized company, improvements are made, and if necessary, the data are reprocessed, etc., before being disclosed through the AI Hub. The open data have a feedback system that continuously manages quality by receiving feedback from the data users and improves data quality in line with AI technology development.

Figure 14: Data Disclosure and Opening Procedures



Source: Authors

c. Data Analysis and Utilization

To help increase the utilization of the AI training datasets opened through the AI Hub, the government asks specialized companies and experts in each domain to use them in actual deep learning, etc., and analyze and take measures to improve aspects such as data quality and model learning effectiveness.

The government also has a separate safe zone for health care data—which contain a lot of sensitive information—so that individuals can analyze and utilize such data without data leaks.

II. Data Platforms

2.1 Open Government Data Platform

2.1.1 Core Assets: Data

Figure 15: Overview of the Public Data Portal



Source: Public Data Portal (data.go.kr)

According to the Act on the Promotion of the Provision and Use of Public Data, all publicly held and managed data that are available to be open should be registered with the Open Government Data Platform (data.go.kr) and disclosed to users. In other words, this platform provides the public with data created, acquired, and managed by public institutions in a single location. The portal satisfies the public's right to know in various easy and convenient ways, such as file data, open API, and visualization. Since the formal establishment of the platform in 2013, the registration and use of public data within it has increased rapidly. As of April 2022, it contained 68,841 datasets provided by 994 institutions.¹⁰

The platform contains a total of 148,853 data entries (based on tables)—including data from 16 categories¹¹ classified according to government work classification standards and national key data in 147 fields. The national key data refers to data that are opened by selecting fields characterized by high effectiveness and urgency, with a focus on demanders (such as citizens and businesses) and refining and processing them into a form that is easy to use for and by the private sector.

10 As a result of proactive efforts to open public data, Korea achieved first place in the OECD public data evaluation three consecutive times (in 2015, 2017, and 2019), and won the Main Prize in the OGP Open Government Awards (December 2021).

11 Education, national territory management, public administration, finance, industrial employment, social welfare, food and health, cultural tourism, healthcare, disaster safety, transportation and logistics, environment and weather, science, agriculture and fisheries, unification and security, law.

There are three main ways¹² through which the data in the Open Government Data Platform are provided: file data, open API, and LOD. File data are provided in the form of data files (csv, etc.), document files (hwp, pdf, etc.), and spreadsheets (xls, etc.) for download. Of the 69,000 cases that are currently provided, 51,000 cases are in the form of file data, which represents a very high proportion. Open API is operated and managed by the organization that opened the API, hence, when the organization's data is updated, it is reflected in real time. For large capacity data that are frequently updated, such as bus operation and weather data, the API method is advantageous. LOD refers to a form that can be used as a single knowledge base by linking the various open data related to each other. LOD is advantageous when there are a lot of highly related partial data, including those regarding people, names of places, and book data.

2.1.2 Economical and Social Assets: User Service

The key function of the Open Government Data Platform is to serve as a point of contact between the private use (demand) of the public data and the opening (supply) of public institutions. It provides a data search function (data search) with an intuitive user interface/ user experience (UI/UX) design to ensure that lay users can easily search for and acquire the data they need and a data map (national data map)¹³ that illustrates the current status of all public data and providers.

Moreover, users can request data that has not yet been disclosed (data requests) from all registered public institutions through the Open Government Data Platform, without having to directly visit the data-retaining institution.

To facilitate the use of public data, the platform also provides a service that processes visualized information to ensure the data can be visually understood and builds an open innovation ecosystem based on the public data by creating a developer network community in which advanced users can participate. Accordingly, platform users directly collect and process the data, or create services using the open data, and a differentiated participatory function is implemented for each user to share the service created as such.

The Open Government Data Platform also offers programs to promote the use of public data. Such programs provide development tools such as the related software and API, a technology support training program that can program the provided data, and an advertisement function that can promote and market the results of use. Consequently, the public can not only achieve the right to know based on data openness but can also strengthen their ability to use the data to generate useful information.

The Open Government Data Platform also provides public-private cooperation services to support companies that use the platform. Its Open Square-D, an offline-oriented public data utilization support space that exists at six locations nationwide (Seoul, Busan, Gangwon, Daejeon, Daegu, and Gwangju) supports the use of public data by start-ups and prospective founders. It provides comprehensive support for people with ideas about public data to gather, exchange experiences and technologies, and grow towards commercialization and entrepreneurship. It supports the entire life cycle of public data utilization, from the realization of ideas based on public data through start-up to sustainable growth. Physically, it provides a collaboration space that anyone can visit freely, a test environment for testing and data analysis necessary for public data service development, and a space founders can occupy to start to commercialize ideas using public data. The space also provides a startup support program (accelerating program), customized consulting (professional matching for public data utilization and start-up support, and data-based consulting), an open data education program (public data-based idea derivation, public data understanding, and practical training operation), and services such as Networking and Demo Day (Data Connection Day, holding

12 Users can download the data or link it with an open API. The data download file format is CSV, JSON, XML, etc., and it is provided in an open format that is not dependent on specific software; at this time, the data file naming rule is "Provider organization name_dataset name_base date.File format (Example: Seoul_public parking lot_20140501.CSV)." Furthermore, in an open API, the provision of data is supported by complying with the attribute information for each item.

13 Strengthens the location identification and related information search by visualizing and showing relevant data in graph form to ensure that the government can search the location of the data held and opened by the government and annual relationship between the data (Step 1: Perform keyword-based search → Step 2: Perform meaning-based search).

seminars on startup-related issues, networking and information sharing), and investment promotion activities (investor relations).

Figure 16: National Data Map

NATIONAL DATA MAP



Source: Public Data Portal (data.go.kr)

2.1.3 Technical Assets: Hardware and Software Infrastructure

To connect and collect the data produced by all Korean public institutions and provide the public with various functions to utilize the data, the system infrastructure of the Open Government Data Platform is based around a core module of various complex functions and services. It has adopted an open standard structure for linking and efficient management of data from each institution, thereby providing common rules for data set management and opening and ensuring inter-operability (universality).

The Open Government Data Platform system configuration consists of a public data provision part, which forms the basis of the platform, and the various service parts using the public data. The public data provision part consists of an open API to ensure that individual public institutions can transmit data to data consumers (individuals, private businesses, administrative institutions, and other public institutions) through a platform via a link. The open API operates using an authentication key-based security structure to ensure data integrity throughout the data delivery process. Based on this, the platform provides users with functions such as data specification, usage information, and API malfunction information. In addition, the platform provides a dedicated interface function for the data provider so that data provision and update, metadata management, and other functions can be performed easily.

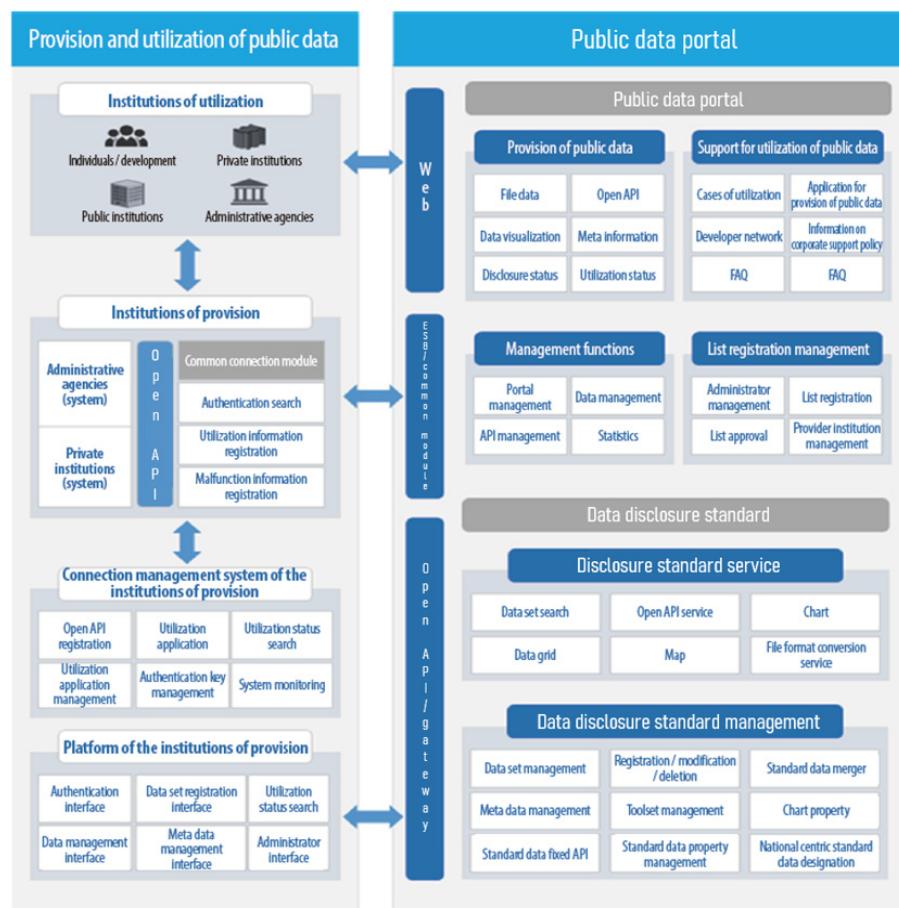
The metadata management of the Open Government Data Platform is a key factor in implementing an integrated management system. For this reason, the government applied the meta management system for the central administrative agencies, local governments, and public institutions to prepare a system for collecting and managing metadata¹⁴ retained and opened by each institution. Beyond this, the government is promoting efforts to standardize and secure the quality of the retained and open data through the standardized management of the metadata based on a standard (data) glossary.

¹⁴ The information expressing the structure, properties, characteristics, and history, etc., of the data.

Second, the system configuration intended to provide various services such as the Open Government Data Platform consists of a vertical structure for each subdivided function by dividing the (user) service function provided by the platform into a data provision area and a utilization support area. To provide this structured service function, each service function element is implemented in a service-oriented architecture. Each service function is modularized to communicate (connect) between services based on web services, using an enterprise service bus (ESB)¹⁵-based structure. To provide versatility for the public data and offer an efficient management system for data from numerous organizations, as well as data inspection and maintenance (typos, dead links, etc.), the API service provision method, metadata management, file format, open standard service for attribute management and data charting, and an open standard data management system are introduced.

Currently, the public data portal is promoting cloud conversion to respond to the increasing amount of data and data demand based on the government's progress in the area of intelligent informatization. In terms of the open API linkage method between the platform and each individual institution that provides data, both the distributed open API linkage (public method) and the central open API linkage (gateway method) are operated. Because the two methods are mixed in the operation of the open API, issues such as inefficiency in API management and failure to response can result. To address such issues, the Open Government Data Platform is preferentially promoting cloud environment conversion for the major open APIs with high private demand.

Figure 17: Configuration of the Public Data Portal System



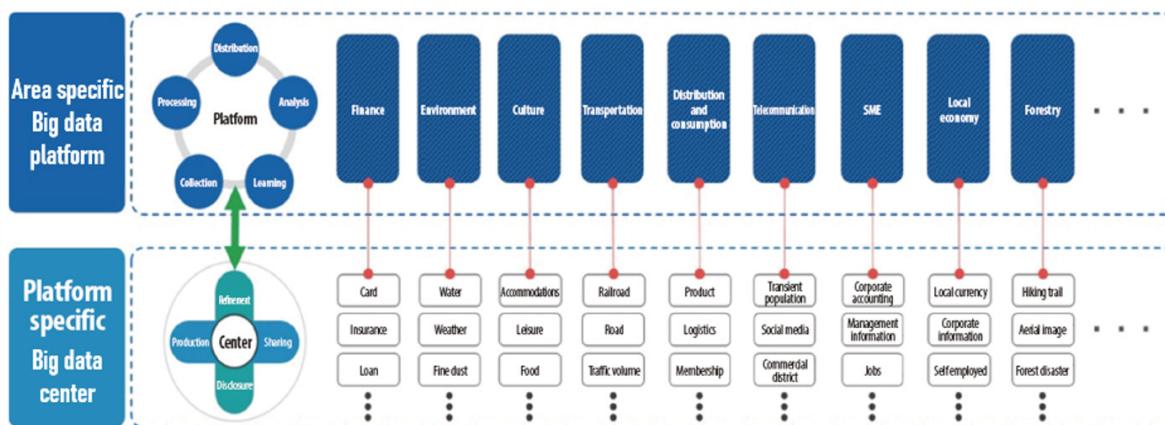
Source: Public Data Portal (data.go.kr)

¹⁵ The enterprise service bus is a service-oriented architecture (SOA)-supported middleware platform for interactions between the software service and application components in complex and diverse system architectures using the bus method.

2.2 Big Data (Market) Platform

2.2.1 Core Assets: Data

Figure 18: Configuration of the Big Data Platforms and Centers



Source: Authors

As of April 2022, big data platforms¹⁶ (150 centers) have been established for 16 industrial fields, including telecommunications, finance, environment, and welfare. The platforms for each field are illustrated in the integrated data map¹⁷ that allows users to easily search and utilize the data from big data platforms in a single location. Currently, each big data platform implements metadata linkage standardization,¹⁸ data quality management, and trading functions to activate data distribution and trading through the platform, and a total of 86,318 datasets and 11,855 pieces of data are distributed.

Table 9: Distribution of Data on 16 Platforms

Area	Data Examples
1. Transportation	Status of prohibited parking areas, private parking lot (status, usage statistics), public transportation usage information, LIDAR (Light Detection and Ranging) collection information, vehicle registration status by manufacturer, car sharing/car-hailing data, on-board diagnostic device data, etc.
2. Finance	Securities company/listed company-related status, listed company disclosure information, ETF/ETN/ELW information, stock lending transaction, fund, bond, e-commerce, bidding/technology evaluation information, insurance company status, insurance fraud, banking transaction, VPN company POS, PG company, online shopping mall data, etc.
3. Culture	Cultural assets' 3D printing design data, cultural facility data/registration status, cultural facility users, film production costs, music video director, game company status, tourist destination/tourism service information, media status, exhibition/education program, major visits and movement data of foreign tourists, etc.

16. Big data platform: provides the foundation for data processing, analysis, and distribution in line with market demand. Big data center: produces and supplies high-value data to the platform.

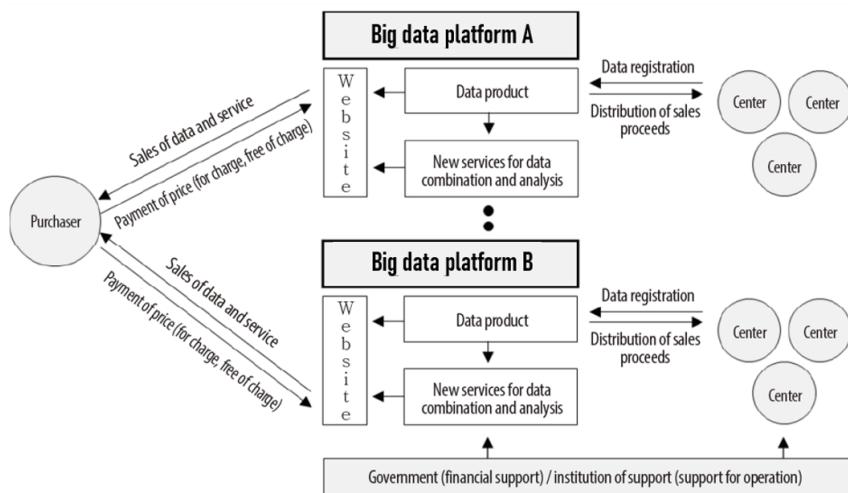
17. The integrated data map (www.bigdata-map.kr) comprehensively demonstrates the data products of the 16 big data platforms, and serves as a portal that makes it easy to find the location of data products through an integrated search.

18. The standardization of platform metadata linkage standards, data standard terms, standard codes, and linkage keys, etc., has been implemented.

Area	Data Examples
4. Forest	Product retail distribution information, smart farm data, areas vulnerable to forest fire information, forest disaster information, hunting prohibited area status, carbon emission, forest product retail distribution information, etc.
5. Distribution and consumption	Domestic corporations' place of business, corporate credit, corporate finance, number of businesses/employees, sales, operating expenses, cost of sales, labor costs, retail/wholesale information, media/content viewing behavior data, etc.
6. SMEs	Startup/closure-related status, welfare information, domestic/overseas investment, M&A, debt status, rehabilitation bankruptcy information, exchange rate, CSV/CSR activity, company certification, support business information, etc.
7. Local economy	Illegal exchange of local currency data, number of use cases by region, monthly order volume, total purchase amount by order product, food delivered by region, location/order product information by region, sales of small business owners, small business data, etc.
8. Communication	Status of wireless mobile operator (MNO), status of specific wireless mobile operator (MVNO), mobile communication station status, frequency allocation status, wireless communication product, wireless mobile operator sales status, wireless communication technology status, nationwide electric vehicle charging station status, charging time by vehicle, real-time charging status of use, etc.
9. Healthcare	Infectious diseases, past chronic diseases, acute myocardial infarction, organ transplantation, immigrant patients, death statistics, medical accident statistics, information by disease type, disease sequelae, injuries, rare intractable disease data, etc.
10. Environment	Analytical diary, air pollution ranking by region, fine dust mask market status, companion animal registration/accident status, coastal/inland wetland information, urban ecology status map, sewage pipe network map, river water quality inspection status, waste disposal status, solar power plant status, fossil fuel consumption, etc.
11. Firefighting and safety	Chemical disaster, smart city, heatwave information, firefighting equipment, number of incidents per hour, emergency rescue calls, possession/utilization of local firefighting resources, special building data, etc.
12. Life log	Citizens' physical condition survey, nursing home/bed status, emergency medical institution status, funeral home status, medical device manufacturer/permission information, medical substance status, treatment prescription amount, environmental harmful factors, chronic disease treatment information, etc.
13. Agriculture and food	Nutrient information by soil type, crop yield, rice production, food safety/food hygiene/food industry, imported food, large-scale retail agriculture and food POS, e-commerce agriculture and food transaction information, livestock, rural commercial area analysis, smart farm production data, etc.
14. Smart security	Judicial precedent information, corporate reputation, crime case, road map image, AI image control object recognition data, unmanned enforcement equipment installation status, traffic analysis, corporate crime analytical information, etc.
15. Marine fisheries	History of accidents within the port, work information in the port, vessel entry/exit, cargo import/export, vessel performance analysis, reef information, aquatic product distribution, real-time terminal import/export, aquaculture, water quality management, mortality, tide data, etc.
16. Digital industry innovation	Status of holding domestic and foreign technology/patents, industrial structure status, market share/order status, employment insurance, stock price information, corporate credit, company certification, startup investment information, reference standards, 3D printer factory, company recruitment, corporate review data, etc.

2.2.2 Economical and Social Assets: User Service

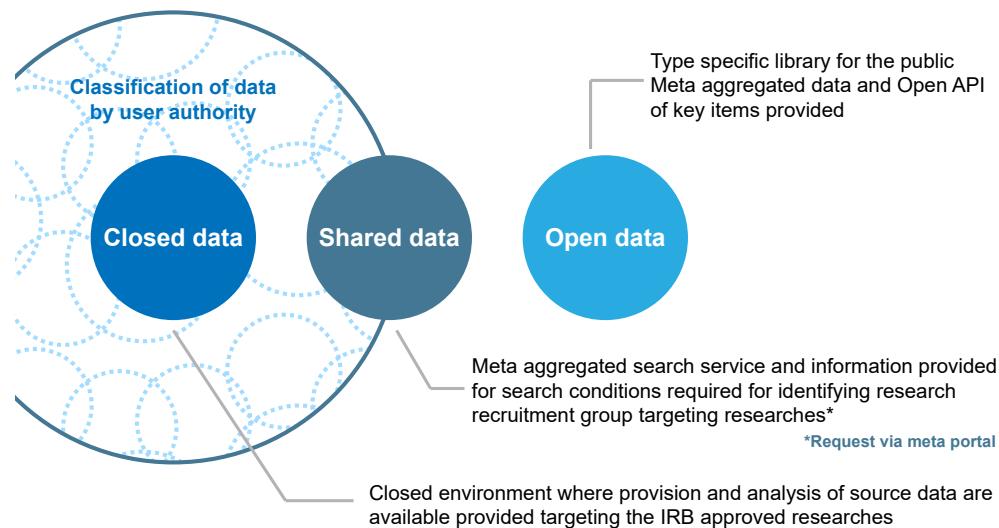
Figure 19: Configuration and Operating System of the Big Data Platform



Source: Authors

The core function of the big data platforms is to create an open trading environment in which consumers can search for and purchase data products. After undergoing a simple membership sign-up process, any data consumer can purchase the data products they seek on the platforms—for a fee or free of charge. However, on some platforms, the membership sign-up terms and conditions and the data usage conditions are restricted due to the characteristics of the data products. For example, on a small business platform, membership is limited to corporate members, whereas on a healthcare platform some raw data are provided only for researchers who have been approved by the Bioethics Committee (IRB).

Figure 20: Permission to Use by Data Type of the Healthcare-Related Big Data Platform

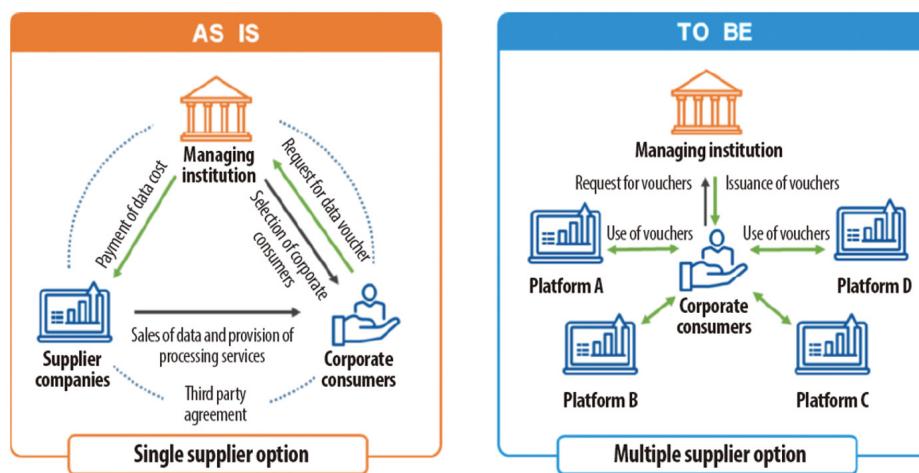


Source: OECD (2022)

The big data platforms also provide the Data 114 service, which connects the user (demand side) who seeks the data and the company (supply side) that seeks to supply the data. Data 114 finds the companies that can provide the users with the data they need or refers the companies that can process and provide such user-customized data.

Moreover, the Trading Support Center has been established for each platform to support data trading and distribution by strengthening the foundation of data trading and distribution such as data standardization, the data pricing model, data de-identification, and legal. In addition, the Voucher One Service has been introduced to ensure that consumers can freely purchase and process data on each data platform. A data-demand forecasting service for the data providers is also provided by analyzing the status of use of the data vouchers and data trading volume of each platform. The big data platforms have also built a trading and analysis integration service (data-as-a-service) environment that utilizes various large-capacity data accumulated on a cloud platform to activate the data trading services, improving convenience for both suppliers and consumers and facilitating the trading market.

Figure 21: Expansion of the Utilization of Data Vouchers



Source: Korea Data Agency (2023)

One of the services that differentiates the big data platforms from the other data platforms is the data combination service. The data combination service is intended to combine the different data scattered across each platform to create new data that is provided for users. Data combination means that the platform does not simply sell the data products but processes and converges the various public and private data retained by the platforms to produce new services and high-quality data. For example, by combining weather data and product purchase data (credit card sales-related data, etc.), “weather-specific shopping trend data” can be created, which in turn can be used to provide weather-customized consulting services for retailers and distributors.

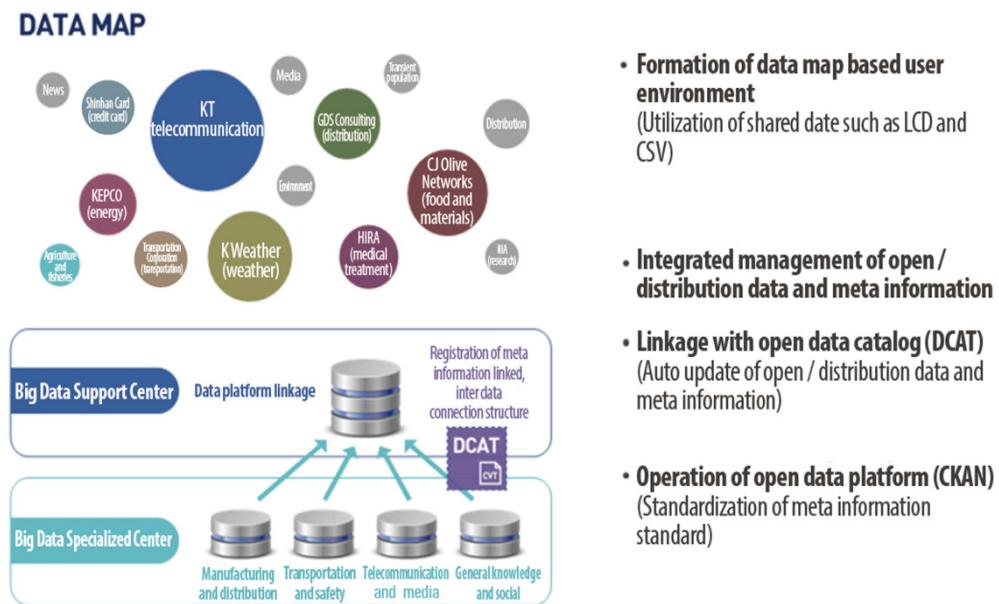
2.2.3 Technical Assets: Hardware and Software Infrastructure

In building the platform infrastructure for the 16 industrial fields and the 150 centers linked to the platforms, the aim is not to apply a single unified system, but rather to adapt to each platform and center according to the nature and uses of the data they collect and process, whereby each platform and center form each system's infrastructure. However, general system configuration guidelines are provided in order to secure flexibility in terms of data use, such as by linking with other platform data and discovering combined services.

The infrastructure configuration of the big data platforms essentially utilizes the private cloud-based common infrastructure to flexibly handle large capacity traffic and related security issues and connect each institution. Of the private cloud services, IaaS (infrastructure as a service) is used in principle, and SaaS (software as a service) allows the participating organizations to autonomously decide whether to use them. Furthermore, while aiming for an open data platform, technical foundations such as CKAN (Harvest),¹⁹ NiFi, and open API are applied, and based on these, the efficient management, search, opening, and distribution of data are supported.

¹⁹ CKAN: Comprehensive Knowledge Archive Network, a web-based platform where open data storage and distribution are available.

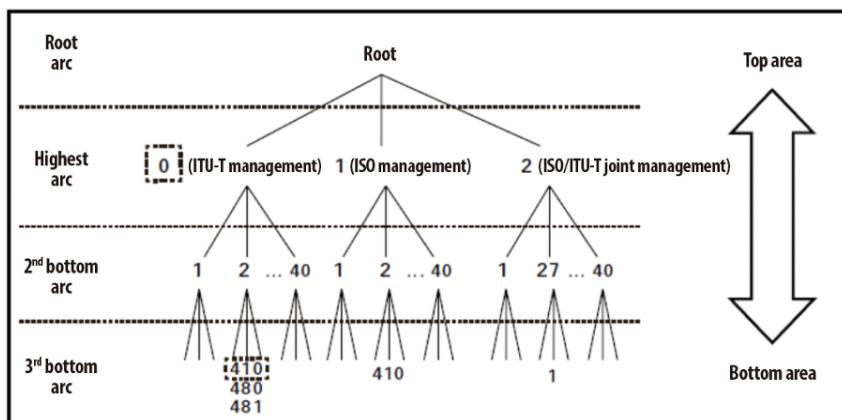
Figure 22: Sharing and Interworking System of the Big Data Platforms



Source: Authors

The big data platforms were built based on a data hub-type model to enhance users' access to the data. The data hub model supports interworking between public and private data open and distribution platforms and the big data platforms by building the metadata that applies DCAT,²⁰ a data catalog standard, along with CKAN for efficient data interworking. Furthermore, the big data platforms support data map development to ensure that open and distributed big data can easily be searched, linked, and utilized (lod, rdf,²¹ etc.) to create a web-based data map usage environment and K-ICT Big Data, and builds the data map by using the metadata collected through the K-ICT Big Data Center and DCAT linkage. One of the important purposes for the infrastructure configuration of the big data platforms is to introduce a management system to secure data reliability (quality, source). By prioritizing continuous quality management such as the extraction of error data, identification and analysis of error items and types, the big data platforms are endeavoring to improve and maintain the quality of data, reduce utilization costs, and increase distribution efficiency through interoperability.

Figure 23: Object Identifier System²²



Source: Zdnet (2021)

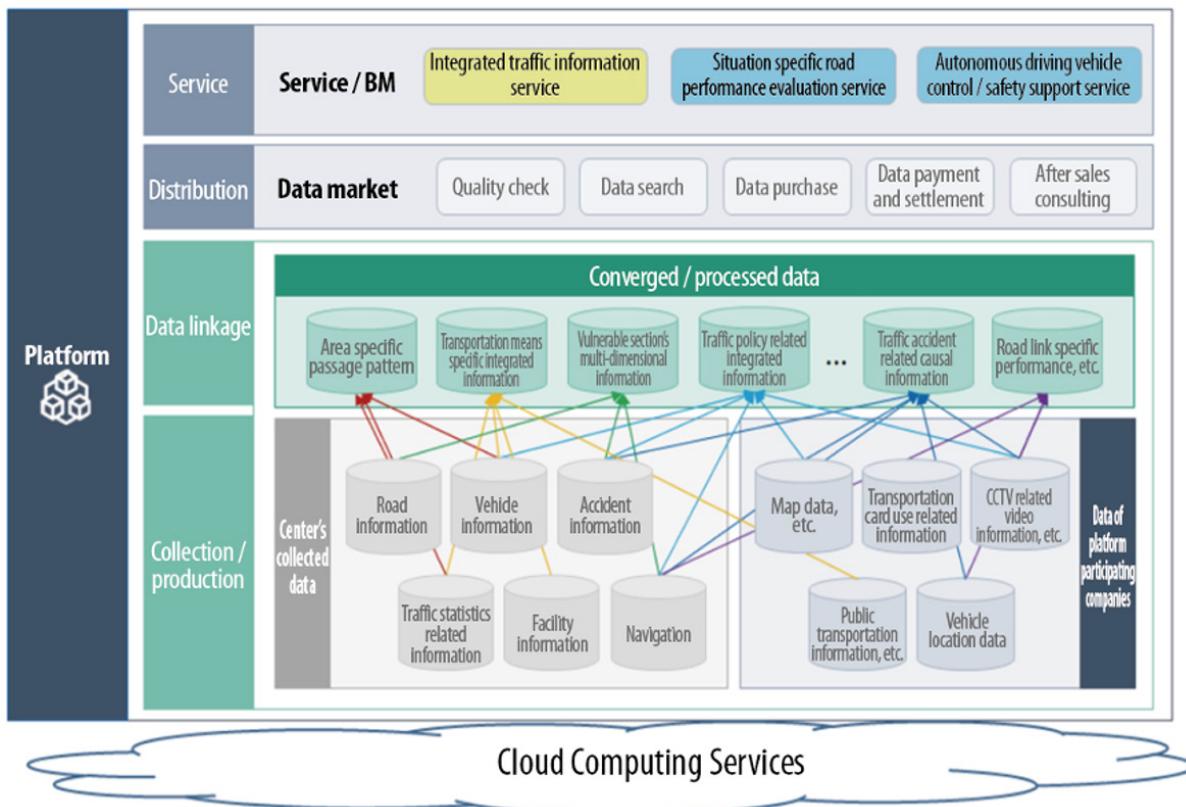
20 DCAT: Data Catalog Vocabulary, a standard for interoperability of the web-based data catalog.

21 RDF: Resource Description Framework, a language for expressing meta information.

22 "Reviewing the 'identification number' system for public-private led big data platforms," Zdnet. (2021). (<https://zdnet.co.kr/view/?no=20210611163147>).

Furthermore, the big data platforms are introducing a technical system to identify the individual data for the statistics and history management of the open and distributed data and to prevent disputes between data providers and consumers. The object identifier (OID, or Object IDentifier),²³ which issues a globally unique identification number, is implemented via the pilot application of some big data platforms' data. Additionally, the big data platforms plan to operate a data identification system and build a system²⁴ for the management of the continuously produced data and the distribution and connection of data.

Figure 24: (Example) Development of a Platform and Center for the Field of Transportation²⁵



Source: UPI News (2019)

- (a) In the traffic big data center, transportation-related data includes data from the expressway (Road Corporation), railroad (Railroad Corporation), regional transportation (Daejeon, Ulsan, Pohang, Jinju), navigation (I-Navi, SKT), floating population (KT), parking (KST Place), and autonomous driving vehicles (Seongnam), etc.
- (b) In the traffic big data platform, the traffic data collected by the center and platform participating organizations are accumulated on the platform, and (i) vehicle traffic and traffic volume, (ii) vehicle GPS, and (iii) traffic data for each section, etc., are linked, producing new data in convergence, including region-specific traffic patterns and transportation mode-specific integrated data.
- (c) Users of the transportation big data platforms can receive the data they require for free and analyze them in the data analysis environment of the platform or in their own development environment.

23 OID is a system developed by international standardization organizations to issue a unique name for a specific object, and a specific OID has a unique value in the world.

24 Implementation of the ORS (OID Resolution System) system development, advancement, optimization, and linkage expansion on a phased basis.

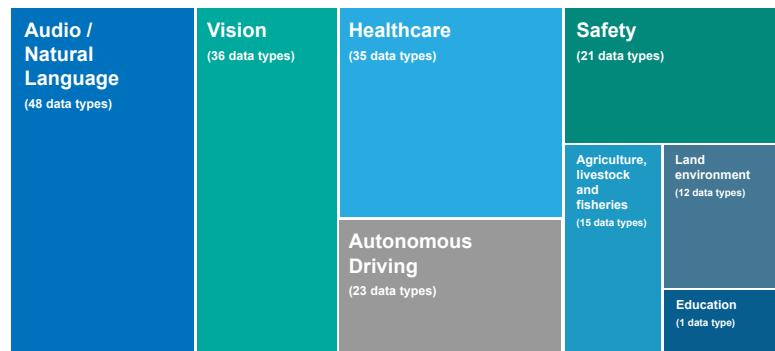
25 "Government builds 10 big data platforms and 100 centers... 151.6 billion won invested over 3 years." UPI News. (2019). (<https://m.upinews.kr/newsView/1065617120735279>).

2.3 AI Hub

2.3.1 Core Assets: Data

The key function of the AI Hub is to collect the source data suitable for AI training and create and disclose the labeling data. As of April 2022, the AI Hub had provided a total of 191 high-quality AI training datasets that are in high demand by industry. There are eight fields of data provided by the AI Hub—vision, voice and natural language, education, land environment, livestock and fisheries, safety, autonomous driving, and healthcare. Furthermore, the data types provided by the AI Hub are divided into image, video, text, audio, 3D, and sensor data. Voice and natural language data (48 types), vision data (36 types), and healthcare (35 types) account for the largest shares. Most of the data sources provided by the AI Hub are datasets created through the AI Training Dataset Project. Among the data accumulated through other government-supported projects, it is providing a dataset that is very useful in developing AI models.

Figure 25: Data Sets Provided by the AI Hub



Source: NIA “AI Hub” (<https://www.aihub.or.kr/>)

organization of the AI Hub, has prepared the standard guidelines for each data development field and introduced several quality verification procedures.

The ultimate goal of the AI Hub is to provide the basic data essential for AI technology development for the public and the private sector, to use them to further expand the data technically, and to open such expanded data for other users again through the AI Hub, thus creating a virtuous cycle ecosystem. The AI Hub also published an authoring tool to produce AI training datasets to ensure that anyone can participate in data labeling. Furthermore, by sharing the AI service model using the data from the AI Hub, AI technology development and service expansion are pursued simultaneously.

2.3.2 Economical and Social Assets: User Service

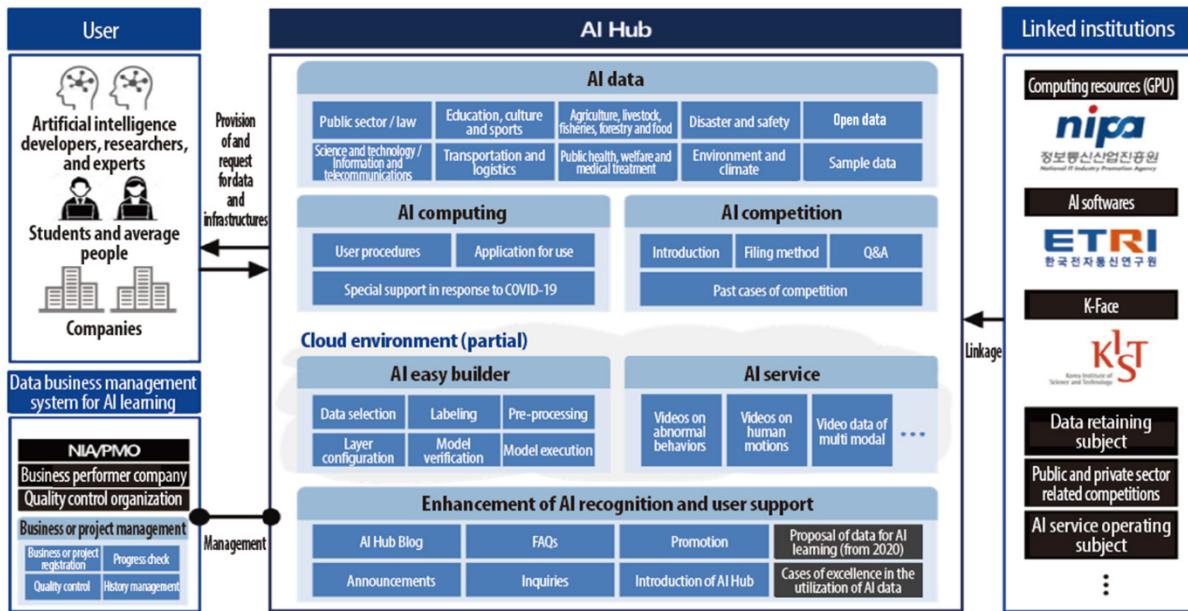
The AI Hub basically provides the AI computing resource support,²⁶ AI voucher support, and AI software support.²⁷ Furthermore, the AI Hub is pursuing the role of an open innovation platform by operating a regular contest (an artificial intelligence playground) that utilizes the core data and by providing developer community functions. In addition, the AI Hub is endeavoring to improve the convenience of searching and using the AI training datasets by promoting linkage with related public and local government portals and data provided by the private sector.

The AI Hub permits the use of data not only for non-profit purposes such as research and development by public institutions and academia, but also for commercial purposes, including AI-related products and services developed by private companies. To this end, AI training datasets that are technologically and industrially promising are being built and opened for broad use for AI application development. The NIA, the dedicated operating

²⁶ Provides support for AI-specialized GPU-based computing resources for processing the large-scale datasets required for AI product and service development for small and medium-sized AI venture companies, universities, and public institutions which require high-performance computing resources.

²⁷ There are four general types of AI software (language processing, voice intelligence, visual intelligence, and conversation processing) which can be provided by the AI Hub, and the AI software provides services via API.

Figure 26: Configuration of the AI Hub Service



Source: Authors

Figure 27: AI Hub's Corporate Support Service

Classification		Details
Secure and Process Data	AI Training Dataset Project	It promotes the revitalization of the AI industry and the creation of high-quality and large-scale jobs by creating and opening AI training datasets.
	Big Data Platform Project	This is designed to engage the public and private sectors to collaboratively build the big data platforms that collect and disclose the data necessary for each industry, and to build the centers that continuously provide data to these platforms.
	Data Voucher Project	It provides companies data vouchers to help create the ecosystem of data demand and supply and to promote use of data in all industries.
	My Data Project	It is a government service in which citizens access, store, and utilize their own information that the government owns such as administration, healthcare, finance, education, and etc. in diverse information systems.
Provide Computing Resources (GPU)	High-performance Computing Resource Supporting Project	It supports AI R&D infrastructure such as computing resources and development environment to improve productivity and strengthen competitiveness of small and medium-sized startups that want to research and develop AI products and services.
	AI Data Safety Zone Project	It provides a physical space with the secured network disconnected from the Internet, high- performance computing power such as GPU and storages, and AI analytic and visualization tools to companies and citizens.
Develop AI Services	AI + X Project	It aims to incorporate AI into all industries and the society as a whole to accelerate innovation and create new markets.
	AI Open Competition Project	It promotes the use and spread of AI by discovering AI related ventures and small and medium-sized companies with outstanding competitiveness in AI.
Secure for Cloud Development	Cloud Flagship Project	It aims to support the development of cloud-based services for companies to accelerate transformation and diffusion towards cloud.
Link Demands	Data/AI Service/Cloud Voucher Project	It is a fund that the government provides companies financial supports when they purchase data, AI services, or cloud services.
	Digital Service Contract Service	It a new procurement system in which the government can purchase an AI related digital service from the digital service catalog without going through a long and complex traditional procurement process.

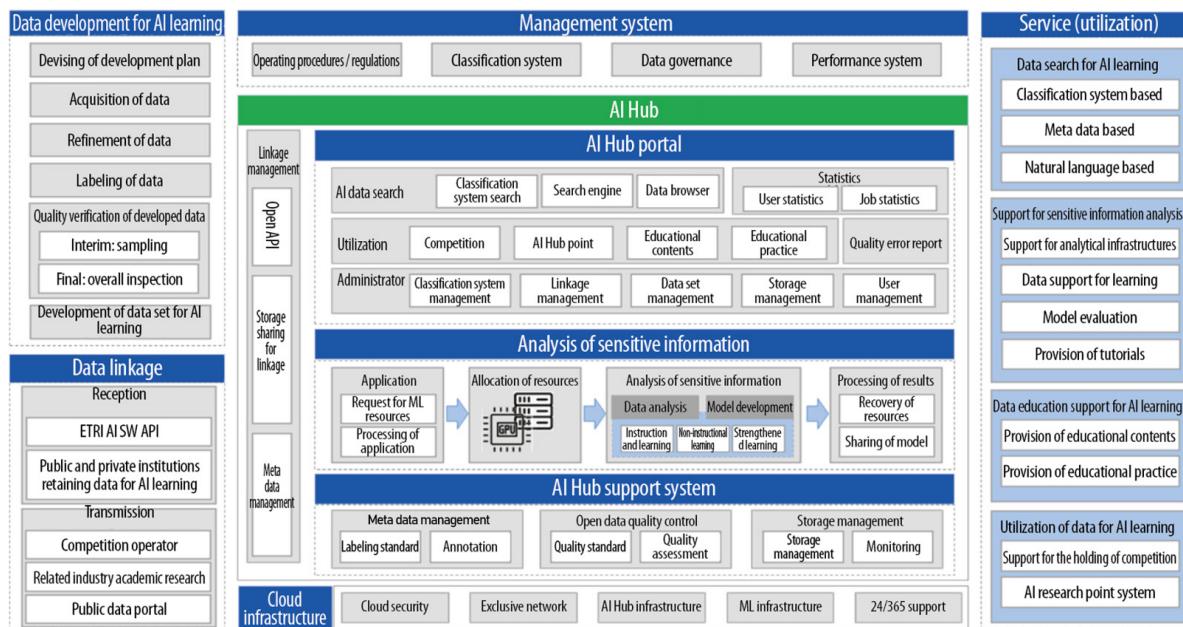
Source: Authors

The AI Hub not only provides AI training datasets, but also discloses the samples of AI application services by using the data and provides methods for using the source code of the AI models used for AI service development, allowing other companies and researchers to add additional data and enabling their use and dissemination. The AI Hub provides a toolkit for creating AI training datasets in order to build the data for each dataset, should the users require additional data processing and labeling work. The AI Hub also performs functions that directly support the business activities of companies. Vouchers are issued for small and medium-sized ventures and mid-sized companies (companies of demand) that need to introduce AI solutions, and the companies of demand can use these vouchers to receive technology development support from the suppliers with technology to introduce the optimal AI solutions. Through the issuance of such vouchers, the AI Hub is contributing to the creation and spread of the overall AI industry ecosystem by providing opportunities for small and medium-sized venture companies (suppliers) that develop artificial intelligence solutions to enter new markets.

2.3.3 Technical Assets: Hardware and Software Infrastructure

The AI Hub has built a cloud computing-based service user environment to ensure that anyone can easily and conveniently access and use the data. It is possible to process the AI training datasets using the cloud computing service provided by the AI Hub without the user directly downloading the data onto their computer. Furthermore, it is possible to use the HW resources for the AI analysis to ensure that the users can directly perform the AI analysis and modeling without purchasing expensive HW equipment (i.e., GPU).

Figure 28: Configuration of the AI Hub System

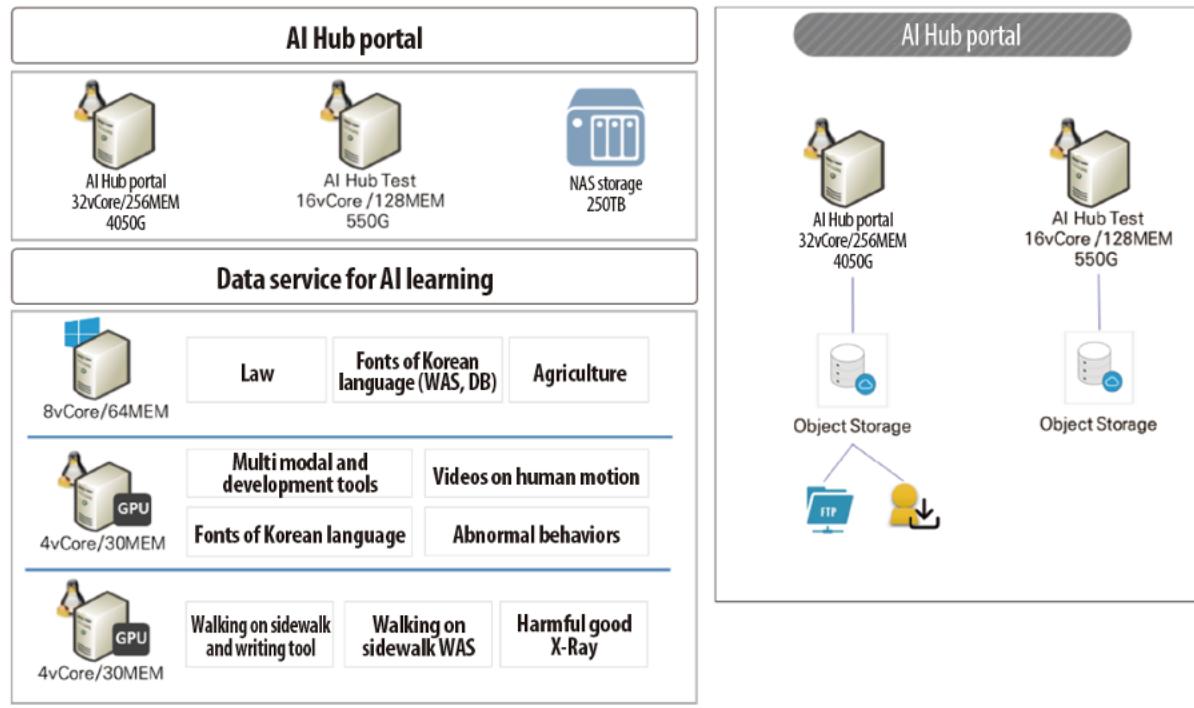


Source: NIA "AI Hub" (<http://aihub.or.kr>)

The AI Hub's cloud-based infrastructure can be divided according to the functions of the AI Hub portal part and the AI training dataset service part. First, the AI Hub portal part is focused on basic service functions such as search and utilization of the AI training datasets and reporting statistics and quality errors. In addition, to secure the safety of the data, including sensitive information, a “sensitive information processing policy” has been established and applied, which provides a basis for data utilization without concerns over data leakage and the infringement of rights. The sub-area of the AI Hub portal part consists of the sensitive information

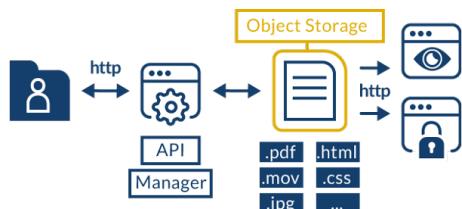
analytical function; the support system area, which performs metadata management, data quality control, and storage management; and data linkage and governance.

Figure 29: Cloud-Based Infrastructure Configuration of AI Hub



Source: NIA “AI Hub” (<http://aihub.or.kr>)

Figure 30: Concept of Object Storage



Source: Authors

and integrates with other services provided within the cloud service along with the linkage function.

Second, the AI training dataset service applies large capacity processing technology and object storage technology to support convenient data search and file access. The object storage technology is excellent for processing unstructured data because it assigns an identifier to each piece of data, stores it in a container, and calls the identifier only when necessary. The object storage technology facilitates access to data by designating identifiers for metadata, redundant data management, and content life cycle management; secures efficiency in unstructured data management;

III. Analyzing Korea's Data Policies and Practices

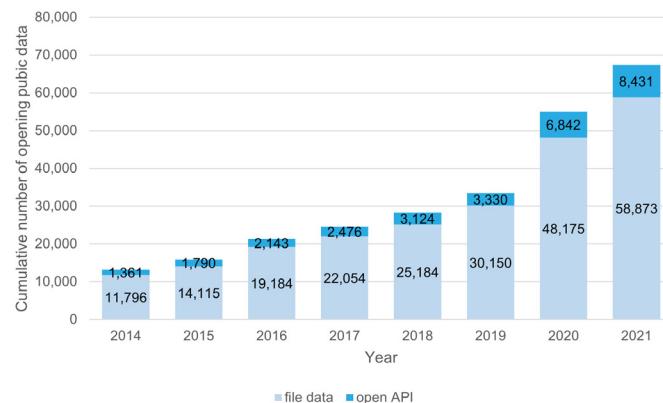
3.1 Performance of the Data Policies

3.1.1 Results of Opening the Public Data

The results of opening the public data has resulted in first, the significant expansion of its content and scope. In 2018, a total public data holding status survey was conducted on all public institutions (about 780) and as a result 28,308 datasets were opened as public data (Figure 31). The cumulative total number of datasets opened as public data continued to increase to 33,480 in 2019, 55,017 in 2020, and 67,304 in 2021. Of all the public data, 492 datasets were opened across 96 fields in the case of the national key data with high socioeconomic ripple effects (i.e., real estate transaction price information, commercial area information, and building information). The cumulative total number of opened datasets as national key data was 33 in 2016, 48 in 2017, 77 in 2018, and 96 in 2019.

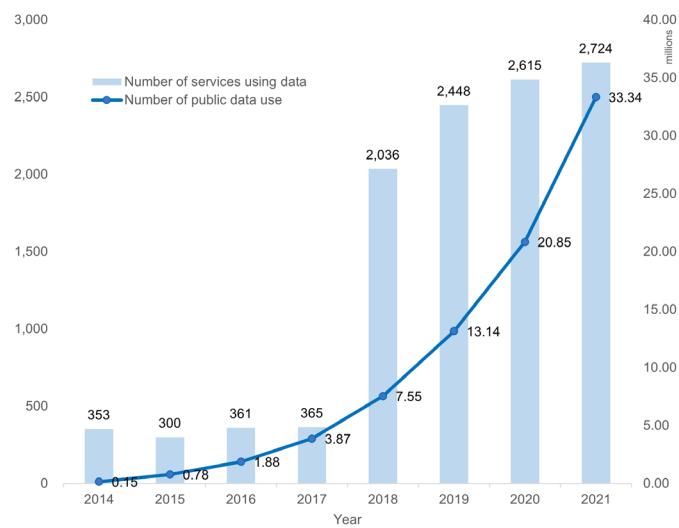
Second, the private use of public data and the revitalization of start-ups has been achieved. As demonstrated in Figure 32, private use of data through the Open Government Data Platform exceeded 3.33 million counts in 2021. The cumulative use of public data was 7.55 million counts in 2018, 13.14 million in 2019, and 20.85 million in 2020. Furthermore, there were over 2,724 cases of private services utilizing public data in 2021. The number of public data use counts in the private sector was 2,036 in 2018, 2,448 in 2019, and 2,615 in 2020. Examples of private use of public data are iamSchool (providing information on school notices such as school announcements and home correspondence) and Jikbang (providing information on real estate sales, lease, and rent including studios, apartments, and condominiums). Furthermore, the revitalization and growth of start-ups using public data was supported. Open Square-D, a space dedicated to utilization of public data and supporting start-ups, began operating in Seoul in 2016, Busan in

Figure 31: Cumulative Number of Opening Public Data (Unit: counts)



Source: NIA (2022). Statistics on Opening the Public Data.

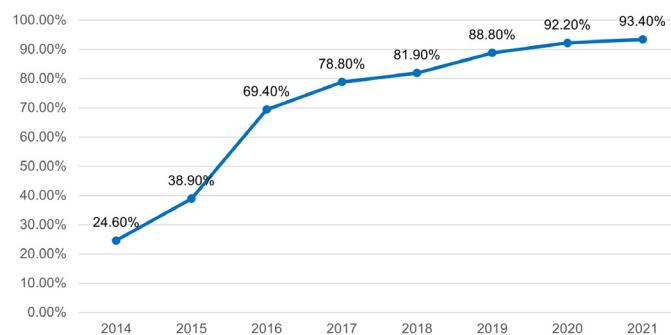
Figure 32: Number of Utilization of Public Data (Unit: counts)



Source: NIA (2022). Statistics on Opening the Public Data.

2017, Gangwon and Daejeon in 2018, and Daegu and Gwangju in 2019, and comprehensive support was provided for 130 projects by 77 companies using public data through the collaboration of 13 specialized agencies.

Figure 33: Trend in the Proportion of Public Data in Open Format

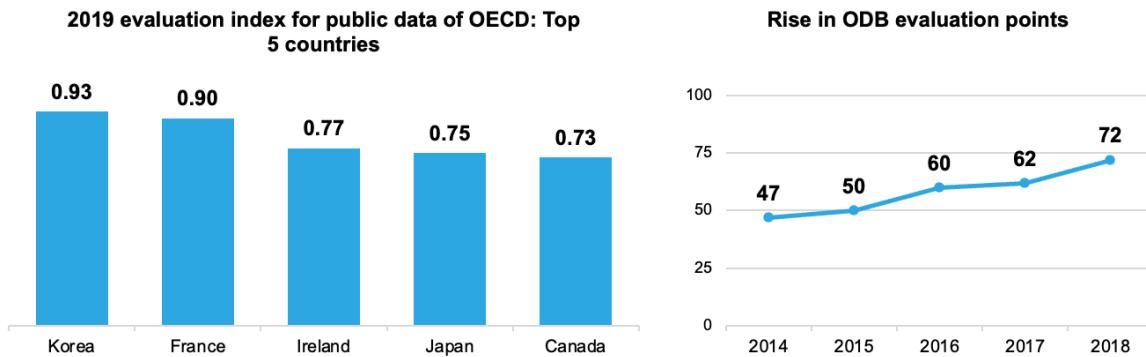


Source: NIA. *Statistics on Opening the Public Data*.

machine-readable open format, having risen from 81.9 percent in 2018 to 88.8 percent in 2019, 92.2 percent in 2020, and 93.4 percent in 2021 (Figure 33).

Fourth, Korea's global rank in terms of public data has risen. Korea was ranked first in the world three consecutive times in the OECD's public data evaluation (2015, 2017, and 2019). In terms of the OECD's public data evaluation index, Korea has a score of 0.93, which is higher than France (0.90), Ireland (0.77), Japan (0.75), and Canada (0.73) (Figure 34). Furthermore, Korea ranked among the world's top four (champion group) in the World Wide Web (WWW) Foundation's Open Data Barometer (ODB) evaluation. Korea has seen a significant rise in its ODB evaluation ranking, from 17th in 2015 to 8th in 2016, 5th in 2017, and 4th in 2018.

Figure 34: Status and Trend Related to the Global Evaluation of Public Data



Source: OECD (2020).

In addition to the quantitative results of data opening, various private sector services have emerged as public data opening has expanded since the outbreak of COVID-19. A typical example is the opening of public data in response to the spread of COVID-19 and the production and spread of COVID-19 maps and related services. Since Korea's first COVID-19 case was verified on January 20, 2020, the Korea Centers for Disease Control and Prevention (KCDC) has released data on COVID-19 (i.e., data on the route of movement of confirmed cases, quarantine places for confirmed cases, number of confirmed cases, number of symptomatic cases, etc.), which was provided in the form of a file or open API through the Open Government Data Platform and its own website.

Third, a data management system such as public data standardization was procured. One hundred and twenty types of opening standards were enacted for the data jointly owned by many organizations with high private demand (i.e., data with high ripple effects when integrated and opened nationwide, such as parking lots, public toilets, and tsunami shelters). The cumulative number of public data open standards rose from 11 in 2014 to 79 in 2016, 109 in 2017, and 120 in 2018. Furthermore, 93 percent of the opened public data is now in

Only a day after the KCDC opened the public data related to COVID-19, the COVID-19 Map emerged. A university student made the COVID-19 Map in a single day using public data on COVID-19 provided by the KCDC on January 30, 2020, and released it free of charge. COVID-19 Map users were able to check on the movement of confirmed cases, which was updated in real time using the confirmed case data provided by the KCDC. The COVID-19 Map had recorded 2.4 million cumulative views by its second day and had 13 million cumulative views and an average of 1 million daily visitors as of February 13, 2020.

Since the emergence of the COVID-19 Map, private sector developers and companies have actively developed and produced services to counter the spread of COVID-19 and track other infectious diseases. Some developers and companies downloaded the public data from KCDC as well as the National Geographic Information Center and the Korea Regional Information Development Institute to create an advanced and detailed COVID-19 Comprehensive Situation Map. This map provides a customized service that can satisfy the goals of different users (e.g., verifying the movement of a confirmed patient, confirming the quarantine location of a confirmed case, and confirming the location of hospitals for diagnosis and treatment). Some developers and companies are using AI and machine learning to develop a COVID-19 tracing dataset and open it, or to provide a visualization service to help the public understand the status of COVID-19 and forecast trends of COVID-19 and other infectious diseases by developing relevant models.

Services such as the COVID-19 Map not only highlight the potential for developing innovative public services through opening data at the national and social levels, but also show how the public and private sectors can produce creative solutions that realize a public benefit. In a sense, the COVID-19 crisis provided an opportunity to confirm the importance of data opening and cooperation in the development of innovative public services. It also helped to raise public awareness and form a consensus between government and the public about the need for and importance of opening public data.

3.1.2 Performance of the Big Data Platforms

As a result of the big data platforms' implementation, 16 big data platforms were developed in the public and private sectors and the data required for the related industries and policy development are being supplied. To seamlessly supply the new data, 108 data centers led by central administrative agencies, local governments, public institutions, and 21 centers led by private companies—for a total of 129 data centers across various fields—are now in operation (Table 10).

Table 10: Centers and Data Products of the 10 Big Data Platforms

Area	Platform operation	Center	Data product
Finance	BC Card	KT, SBCN, Nota, Nielsen Korea, Daum Soft, Mango Plate, Habit Factory, Kiwoong Information & Communication Co., Ltd., Korea Appraisers Association, Korea Financial Solutions (10)	Financial (loan, insurance, securities) and non-financial (telecommunications, SNS, distribution, media, commercial area) data Small business owners' start-up guidance service and national financial life planner service, etc.
Environment	Korea Water Resources Corporation	GDS Consulting Group, National Institute of Ecology, Green Ecos, Korea Institute of Meteorological Industry and Technology, Irexnet, Geologic Resources Research Institute, Environmental Policy Evaluation Institute, Korea Advanced Institute of Science and Technology, Soonchunhyang University Industry-Academic Cooperation Foundation, Novacos (10)	Data such as water, weather/climate, fine dust, geology/disaster, ecology/resources, chemicals/substances, and environmental social media Customized water quality information service and air quality outdoor activity recommendation service, etc.

Area	Platform operation	Center	Data product
Culture	Korea Culture Information Service Agency	National Library of Korea, National Sports Promotion Agency, Data Marketing Korea, Red Tie, Red Table, Culture and Arts Committee, Busan Information Industry Promotion Agency, Yanolja, Korea Youth Promotion Agency, One2CM Co., Ltd., TNS Co., Ltd. (11)	Data such as culture, lodging, leisure, food, commercial area, and book/publishing Comprehensive cultural leisure information service and Korean Wave related business matching service, etc.
Transportation	Korea Transport Institute	KST Place, KT, SK Telecom, Jinju City Hall, I-Navi Systems, Ulsan Information Industry Promotion Agency, Korea Credit Bureau, Pohang Techno Park, Seongnam City Hall, Daejeon City Hall (10)	Data such as real-time traffic volume, public transportation, train, highway, navigation, black box, floating population, and parking Road and public transport improvement services and smart city support services, etc.
Healthcare	National Cancer Center	Sungkyunkwan University Industry-Academic Cooperation Foundation, Konyang University Hospital, Daegu Catholic University Hospital, Yonsei University Industry-Academia Cooperation Foundation, Seoul National University Bundang Hospital, Chonbuk National University Hospital, Seoul National University Industry-Academia Cooperation Foundation, Hwasun Chonnam National University Hospital, Ajou University Industry-Academia Cooperation Foundation, Gil Medical Foundation (10)	Clinical data on 10 most common cancer types (data currently provided on uterine cancer, thyroid cancer, and ovarian cancer) Cancer diagnosis and treatment decision making, anticancer drug research and development, etc.
Distribution and consumption	Maeil Broadcasting Network	Nice D&R, Daumsoft, Dable, Low Flat, Built-On, Shikshin, Onnuri HNC, Korea Postal Service, Jin Plus, Korea Credit Bureau (10)	Data such as distribution of products, card payment, courier invoice, communication, real estate, commercial area, logistics, restaurant, used car price, and social media Preferred restaurant service by lifestyle and online product purchase information service by region, etc.
Communication	KT	Kyonggi University Industry-Academia Cooperation Foundation, Ness, Next Easy, Korea Smart Grid Project Group, Doing Lab, BC Card, Small Business Association, Amazing Food Solution, Open Mate, Incheon Technopark, Korea Internet & Security Agency, The Big Nanum MTN, Conan Technology, Zero to One Partners (14)	Data such as population distribution, commercial district, card use, tourism, transportation card information, and social media Commercial area analysis service and daily living population analysis service, etc.
SME	Duzon Bizon	NICE, Big Value, Korea Industrial Technology Association, Wisenut, Incruit, Korea Trade Information and Communication, Korea Productivity Center, Hanwha General Insurance, Green Technology Center, Sundo Soft (10)	Data such as SME accounting information, real estate, insurance contracts, corporate employment/welfare benefits, and social media Business management information analysis service and job demand forecasting service, etc.

Area	Platform operation	Center	Data product
Local economy	Gyeonggi Provincial Government	Gyeonggi Credit Guarantee Foundation, Gyeonggi Jobs Foundation, Gyeonggi Content Agency, Korea Research Institute for Human Settlements, The IMC, Hanyang University Industry-University Cooperation Foundation, Korea Enterprise Data, Korea Institute of Industrial Technology (8)	Local currency payment and settlement information, corporate information, jobs, credit evaluation, card company information, and data such as Gyeonggi-do population, housing and environment Regional consumption pattern analysis service and customized job matching service, etc.
Forest	Korea Forestry Promotion Institute	Woorim NR, Beagle, Korea Forest Welfare Promotion Agency, Sanya Hanging Industry, Sunsun IT, Aro Information Technology, Woolim Infotech, Market Link, Korea Institute of Oriental Medicine, Infoboss (10)	Data such as forestry, hiking trails, forest trails, bicycles, public transportation, mountain weather, forest disasters, and aerial images Tracking service and forest disaster prediction service, etc.
Fire Safety	National Fire Agency	Sejong Fire Department, JeonBuk Fire Department, Ulsan Fire Department, Jeju Fire Safety Department, Korea Fire Institute, Korean Fire Protection Association, Allitelife, Updater, Korea Fire Protection UBIS Co., Ltd., Prognosis & Diagnostics Technologies (10)	Data such as fire safety, fire industry, disaster insurance, fire IoT, fire safety capability, fire risk analysis, and underground piping safety SAFETY 119, fire association analysis, spatial data visualization, etc.
Smart Policing	Police Science Institute	Suwon City Hall, Korea Land and Housing Corporation, Seoul Credit Guarantee Foundation, S2W, Thecheat, Jirandata, Amgine Securus, e2on, iTRO, Yonhapnews (10)	Data such as incidents by office, safety-related surveys, illegal advertisements and malicious code URLs, natural disasters, and location information of administrative and legal dongs nationwide Providing services such as safety guidance map and prevention of financial fraud
Oceans and Fisheries	Korea Maritime Institute	KESTI, Billion21, Mokpo University, GIST, UNIST, Lab021, HAEWOO Co., Ltd., Shipping & Port Logistics Information Association, Korea Maritime & Ocean University, KOMSA, All Sea Data (11)	Data such as marine industry, marine environment, international cooperation, marine fisheries, fishery resources, fishing village farming, shipping logistics, port operation, and maritime safety Providing services such as vehicle turnaround analysis information, marine environment data statistics, average unit price of imported seafood, marine area information, and impact on shipbuilding

Area	Platform operation	Center	Data product
AgroFood	Korea Agro-Fisheries & Food Trade Corporation	Ezfarm, Ulogisnet Co., Ltd. Jangbogofoodbank, Nicezinidata, Kplus, EZHLD, Korea Trade Statistics Promotion Institute, Jeonnam Information & Culture Industry Promotion Agency, One Data Technology (9)	Data such as recipes, school meal delivery, restaurant consumption patterns, real-time auction data in the national public wholesale market, and reports on trends in the agricultural and livestock overseas market Comparison of wholesale market prices across the country, logistics information of agricultural products, recommendation of shipments by agricultural product, self-diagnosis of pig farmers, and shipment of agricultural products
Life-log	Wonju Severance Christian Hospital	Kangwon University-Industry Cooperation Foundation, Korea University Medicine, Goodoc, Korea Hearing Big-data Center, Bagel Labs, i-sens, K-Weather, Hallym University Medical Center, HealthMax, Huraypositive, Health Bridge, LG Uplus (12)	Data such as life log by disease, life pattern life log, medical measurement life log, and diet life log Provides services such as measuring blood sugar health, measuring cardiovascular and diabetes-accompanying diseases, and measuring electrocardiograms
Digital Industrial Innovation	Korea Testing Laboratory	Korea Enterprise Data, Korea Industry Intelligentization Association, Korea Productivity Center, Korea Insight Institute, FnGuide, GAION, Korea M&A Exchange, HelloDD, Alicorn, STHIS, E&C GLS, Hebronstar, PatentPia, BAX Intelligence (14)	Data such as M&A information, science and technology, finance, corporate information, logistics, import and export, test certification, distribution, certification, finance, investment attraction, patents, and funds Providing services such as connecting data talent sellers and buyers, and challenges for solving industry problems

Source: Big Data Map (www.bigdata-map.kr)

The key achievements of the big data platforms are as follows. First, they support data analysis. Some platforms (56 platforms or 43.4 percent) support a data analysis environment to ensure that users can easily utilize the data accumulated on the platforms. These platforms provide online and offline analysis tools including Python, R, and AutoML, or platforms or analysis companies provide data analysis services for the users.

Second, they allow for data trading. A distribution channel in the form of an open market is created with a focus placed on the big data platform. Data trading has traditionally been focused on specific areas in which the use of data—including financial and corporate information, marketing, advertising, and retail information—directly leads to monetary value. This can diversify business models through the collection and processing of customized data or the development and selling of data analysis models that companies want.

Third, they lay the foundation for data distribution. By building big data platforms based on public-private collaboration and supporting transaction systems, the revitalization of transactions focused on the data platforms is induced. The Integrated Data Map service, which allows users to search the Big Data Platform data from a single location, was launched (March 2020), and the search was linked with the AI Hub, the Data Store, Trade Big Data (February

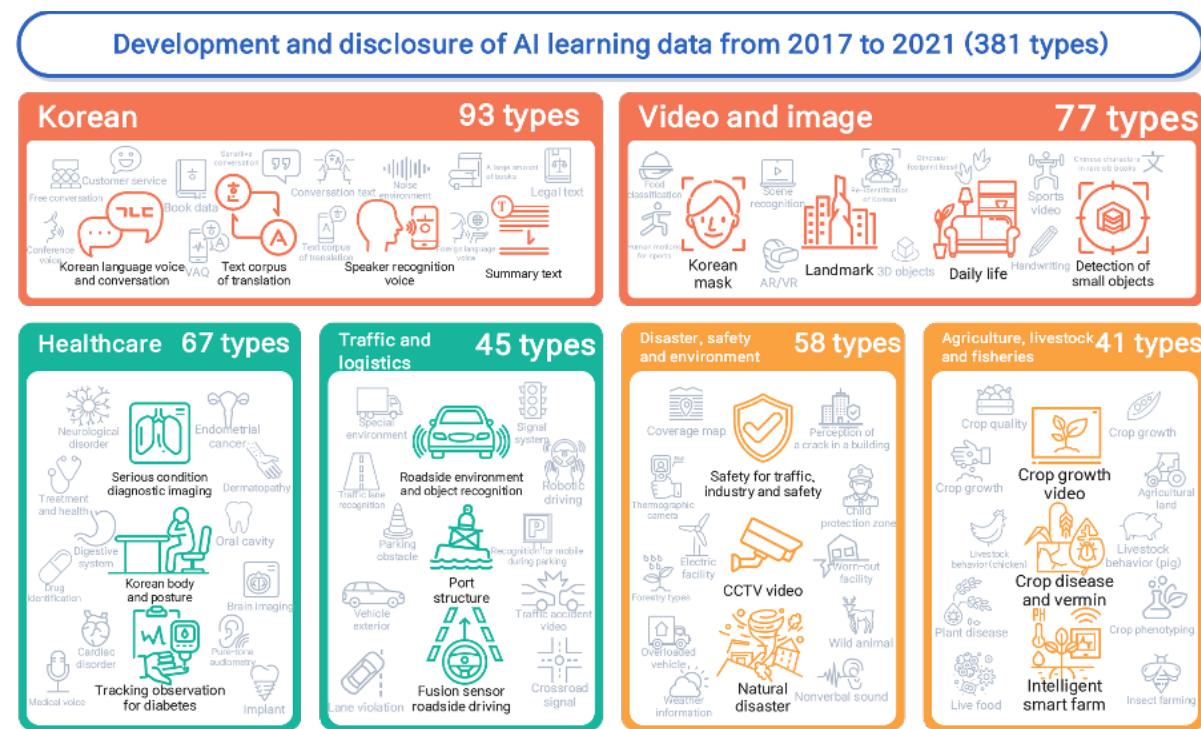
2021), and the Korea Tourism Data Lab (April 2021), while linkage with other platforms was also expanding. To prepare this distribution base, guidelines (of three types) were developed (December 2020) that provide specific information in the fields of price, quality, and law, which data sellers and buyers can refer to when conducting their trade.

Fourth, they promote standardization and quality improvement. Focused on the Open Government Data Platform and the big data platforms, standardization and quality control are promoted for the easy connection and utilization of data. The big data platforms produced standardization and quality control guides from the data development stage (2019) and now provide the data for each platform, which are applied for development and management. Furthermore, based on the common standard terms for public data, a dictionary of related standards was prepared by organizing the public and private terms for each industry required for the big data platforms.

3.1.3 Performance of the AI Hub

Based on the development of the AI Hub, 381 types of AI learning data have been opened (as of June 2021) across various fields. Specifically, the 46.5 million cases of 21 types (including laws, patents, Korean voice, and abnormal behavior CCTV) by 2019, 480 million cases of 170 types in 2020, and 580 million cases of 190 types in 2021 were constructed and opened. The cases were opened to the AI Hub after quality verification and error/validation inspections were completed through an advance opening for developers and experts (Figure 35).

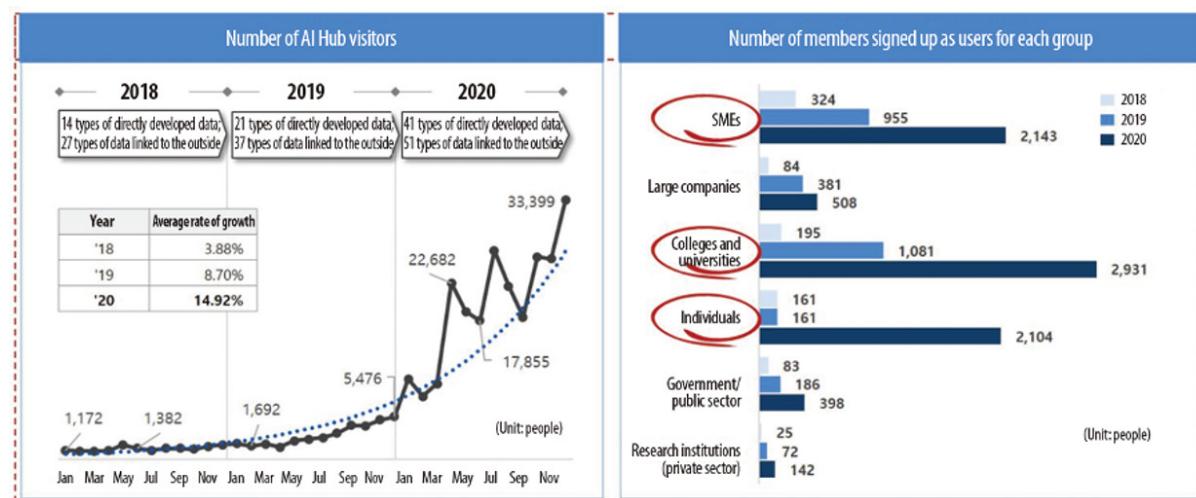
Figure 35: Status of Data Development for AI Learning²⁸



Source: Ministry of Science and ICT and NIA (2022).

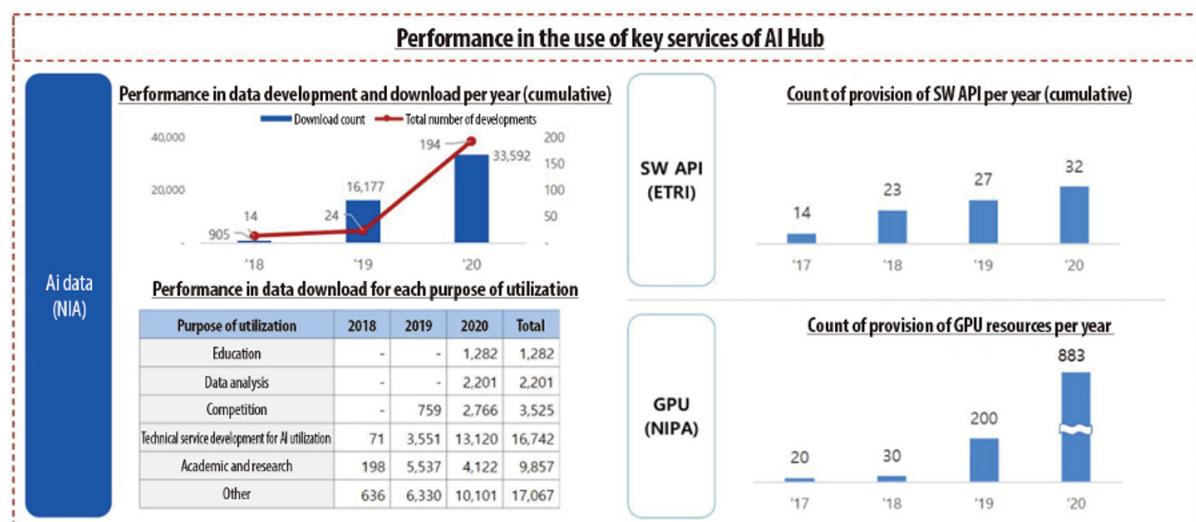
As such, the AI Hub development, AI training data sets development, and the private sector provision achieved the following tangible and intangible results. First, the interest in AI data is growing. Specifically, the number of visitors has increased rapidly in recent years, and in particular, it was found that the interest among small and medium-sized enterprises and startups and students and individuals with relatively low access to AI resources is high (Figure 36).

²⁸ Ministry of Science and ICT and the National Information Society Agency. (2022). "The Open Dataset Project"

Figure 36: Number of AI Hub Visitors and Number of User Membership Sign-Ups

Source: NIA (2021). Statistics on the Use of AI Hub and Internal Data.

Second, the performance in terms of data downloaded has significantly increased since the opening at the end of 2017. The download count was 33,592 in 2020 and had reached 130,901 (cumulative) by 2021. The scope of data utilization is also diversifying from the initial technology and service development and academic research to education and participation in competitions. The number of infrastructure provisions related to the use of artificial intelligence has been increasing. For example, the number of software APIs provided by year and the number of graphics processing unit (GPU) resources provided by year are significantly increasing (Figure 37).

Figure 37: Performance in the Use of AI Hub's Key Services

Source: NIA (2021). Statistics on the Use of AI Hub and Internal Data.

Third, the development and spread of AI innovation services began in earnest based on the development of the AI training datasets and their opening for the private sector. As of December 2021, 110 cases of AI service development and advancement and 33 cases of AI service spread were achieved through the use of AI training datasets. Domestic and foreign academic research and registration of intellectual property rights were also activated, with 79 studies conducted, and 65 instances of intellectual property rights registered. There have

also been achievements from an economic perspective. A large-scale workforce, including young women and women who took a career break due to childbirth, is participating in data development, including work for major domestic artificial intelligence companies, local governments, universities, and hospitals. In 2020, 525 companies and organizations and 40,165 people participated, and in 2021, 552 companies and organizations and 42,917 people participated, which exceeded the target (Table 11).

Table 11: Status of Workers Participating in Data for AI Learning (Unit: number of people)

Classification	Target for 2020	Performance for 2020			Target for 2021	Performance for 2021		
		Direct employment	Crowd sourcing	Total		Direct employment	Crowd sourcing	Total
Current status	20,197	8,770 (21.8%)	31,395 (78.2%)	40,165	25,581	7,811 (18.2%)	35,106 (81.8%)	42,917

Source: NIA (2021).

3.2 Lessons Learned from Korea's Data Policies

3.2.1 Characteristics of Korea's Data Policy

There are three distinct characteristics of Korea's data policy. First, the policy is government-initiated. In the field of data policy, the Korean government is not acting as a judge or regulator, but rather as a player that creates, supports, and leads the entire data cycle. In terms of data collection and generation, the government not only purchases externally produced data and opens the data it already has to the private sector, but also directly builds and distributes the data, as seen in the AI learning data development project. In addition to building the data platforms such as the public data open portals, big data platforms, and the AI Hub for data distribution, analysis, and utilization, private companies participate in platform development to activate data trading and create data markets while pursuing a policy of making data available to the platform. Furthermore, the government is endeavoring to establish various institutional devices to create and revitalize the data market, such as data asset protection, data valuation and quality control, and data trading companies. To raise the level of data utilization by the private sector and society to a larger scale, various institutions have been established to support data utilization, such as the Public Data Utilization Support Center, and efforts are being made to increase the number of data-related employees and establish a related education system.

Second, the policy calls for risk-taking. Because the data have the characteristics of a public good, it is not easy to activate transactions in the market. There is a high potential that the data cannot be sufficiently traded in the market, as it is difficult to evaluate their value before analysis and utilization. Therefore, companies that seek to maximize profits cannot significantly increase their investment in data development and utilization. This leads to a market failure, where the data-related investment does not reach a socially desirable level. Even in Korea, despite long discussions of the transition towards the data economy, the level of data utilization was not high, delaying the transition to the digital economy. Accordingly, strong arguments emerged that it was necessary to raise the availability level to select data with a high utility and a high combined value, and to improve data maturity through a proactive data policy to enhance the understanding and utilization of data by market participants.

Accordingly, the Korean government took on the social risks required to advance the data economy through various platforms and policies to promote data development and utilization—and finance them—while creating opportunities related to data development and utilization through public investment, planning cooperation with various stakeholders,

fostering uncertain early projects and data-related industries, and supervising and supporting the commercialization of various data-related services and products. The government's social burden and the provision of data as a public good not only helped reduce the risks and costs that private companies had to bear while conducting business, but also contributed to the development of the data industry and creation of opportunities for the digital economy. This, again, means that close cooperation and partnership between the government and the private sector has provided the basis for the promotion of data policy.

Third, the policy is social and industrial innovation oriented. When Korea's data policy began in earnest in 2010, the government faced challenges from various internal and external sources. On the one hand, it was a time when the Korean economy needed to take a leap towards a digital economy amidst the development and increased use of intelligent information technology. Then, in the face of a historic severe economic downturn and job shock due to COVID-19, the data policy was promoted as part of the national development strategy to overcome the crisis. More broadly, the data policy was also intended to "design the next 100 years of Korea," by fundamentally changing the country "from an economy that follows to a leading economy, from a carbon-dependent economy to a low-carbon economy, and from an unequal society to an inclusive society."²⁹ In conclusion, Korea's data policy does not stop at the small goal of improving the level of data development and utilization but takes on the nature of a policy that aims to innovate and take a leap forward for Korean society and industry, while overcoming economic and social crises such as COVID-19.

3.2.2 Success Factors of Korea's Data Policy

The first of the success factors of Korea's data policy is the support it receives from national leadership. A data policy involves a great deal of time and many tangible and intangible resources. If it is to achieve its goals, the incentive system for data-related companies in the private sector as well as various legal systems and institutional foundations in the public sector must be improved and reconstructed. This requires strong leadership that is willing to continuously pursue policy and a strong policy governance system that can mobilize a variety of resources.

Since the 1970s, Korea has continuously promoted national informatization and the realization of e-government by making these goals part of the presidential agenda and providing financial and institutional support accordingly. The data policy, which has been promoted in earnest since the 2010s, was also adopted among the presidential tasks. Data policies were put in place such as the Data Economy Declaration of August 31 (2018), the AI National Strategy (2019), and the Digital New Deal Declaration (2020), supported by the president's leadership. Given the status of data policy as a presidential task, aggressive and ongoing fiscal investments to promote the data policy and reorganization of relevant legal systems for the data policy and industrial development were able to proceed smoothly.

The second success factor is the establishment of capabilities to implement data policy. To successfully implement data policies, national leadership must support the governance body of data policy to promote coordination and cooperation between stakeholders in the public and private sectors, mobilize technical expertise, and secure administrative and financial resources across various areas. Also required are capabilities in planning in line with the new policy environment and policy goals, prompt and reasonable decision making, and evaluation and feedback to evaluate the results of project implementation and revise existing policies to fit the new policy process.

The Korean government assigned MSIT³⁰, which has promoted national informatization since the 1990s as the overall data policy managing ministry to secure policy expertise and continuity and to coordinate with various stakeholders in the public and private sectors while proceeding with cooperation. Furthermore, the support institutions of specialized technologies such as the NIA, the Korea Internet & Security Agency (KISA), the Korea Information Society

29 Address by President Moon Jae-in at Opening of 21st National Assembly (July 16, 2020). (<http://webarchives.pa.go.kr/19th/english.president.go.kr/BriefingSpeeches/Others/851>).

30 Previously, Ministry of Science, ICT and Future Planning

Development Institute (KISDI), the Korea Local Information Research & Development Institute (KLID), the National IT Industry Promotion Agency (NIPA), and the Electronics and Telecommunications Research Institute (ETRI) were established to actively respond to technical issues in implementing data policy.

The third success factor is the maintenance of institutional infrastructure such as laws and systems. Reorganization of the legal system is essential to successfully implement data policy. Various legal issues related to the protection and use of personal information were raised when the data policy began to be pursued in earnest. To address these issues, the Personal Information Protection Act, the Information and Communications Network Act and the Credit Information Act (collectively referred to as the “Three Acts on Data Policy”) were amended in 2020. Furthermore, the existing laws related to national informatization and e-government were also amended to enhance the level of data utilization. The Act on Promotion of the Provision and Use of Public Data (2013), the Data-Based Administration Act (2020), and the Data Industry Framework Act (2022) provided a cornerstone for addressing systemic obstacles to the implementation of the data policy.

The fourth success factor is the utilization of digital and human resources previously accumulated in the process of national informatization and e-government. As noted in Part I, from the 1980s through the 2000s, Korea promoted the National Infrastructure Network Project, the National Database Development Project, and the e-Government Project. In this process, digital resources such as a pan-governmental e-government system that would have accumulated the basic data for each sector of national society were developed. The accumulated digital resources operated as a technological foundation to facilitate the transition towards a new data policy according to the changes in the intelligent information technology environment.

Moreover, personnel with diverse policy experiences and expertise were secured in the process of promoting the informatization of national society starting in the 1980s. Such personnel formed a policy network encompassing industry, academia, research, and government, thereby providing new policy ideas in line with the emerging development of intelligent information technology and the data environment, and by participating in various government committees for data policy deployed their expertise and experience into the decision making and execution process. These activities of the expert policy network contributed to enhancing the validity of data policies and the speed of policy execution.

3.3 Challenges in Korea’s Data Practices

3.3.1 Data Collection and Areas of Generation

Notwithstanding the achievements and success factors mentioned above, there are still many challenges to overcome. Although there have been achievements at the level of individual business units, the extent of data opening and utilization is still not high relative to that of the leading data countries. Although there has been collaboration among institutions within the public sector and the aim is data sharing across departments, there remains a large division between the platforms and a silo effect. Furthermore, data utilization capacity within the public sector is still not considered significant, and there is large variation among institutions. Specifically, the following issues were highlighted.

First, in the area of data collection and generation, there is still a lack of quality data. The government and public institutions lack the support and benefits necessary to produce and provide valuable data, and as such, are passive in providing data that is customized for consumers. From the perspective of companies, the opportunities and incentives to create new data-based businesses and innovation through the discovery and distribution of valuable data are inadequate.

Second, there is a gap between data production and collection capabilities and there is a data shortage by sector and company. In certain fields, such as finance and communications, data production and collection has mainly been focused on large companies and the need to develop various data by field is increasing, with the difference related to the collection gap.

3.3.2 Aspects of Data Distribution and Trading

First, uncertainties in demand and a lack of distribution channels are limiting the vitalization of data trading. Data supply companies are reluctant to invest in data production, collection, and processing for sales because they do not know the exact market demand. Data sellers prefer to sell through data platforms but are aware that the data platform-based distribution channels are inadequate³¹. This mismatch between supply and demand is acting as a constraint that prevents the activation of data trading.

Second, it is difficult to identify the source of data, resulting in very high data trading costs. The emergence of various data platforms paradoxically leads to the fragmentation of distribution channels, thereby increasing the data-related transaction costs desired by the users. Because the metadata format varies for each data platform and important items are omitted, it is necessary to manage the integrated data-related metadata.

Third, there are restrictions on trade negotiations due to information asymmetry, such as standardization, quality, and price calculations. If different standardization and quality controls are applied for each data platform, it may be difficult for users to rely on the data products. For example, even in areas such as transportation where high-level raw data is secured, there is a risk of differences in standardization and quality format because the purpose of collection is different. Given the nature of data, it is difficult to determine their value until the data are analyzed, and as such, trading on the platform is restricted.

3.3.3 Aspects of Data Analysis and Utilization

First, there is a gap between the different companies and institutions in terms of their data analysis and utilization capabilities. Unlike large-scale companies, institutions, and data-specialized enterprises, SMEs and general companies lack the capability to utilize data, hence the use of data across industries and society is somewhat slow. In 2020, the rate of adoption of big data by companies was 13.4 percent but was 35 percent among companies with more than \$78.56 million in annual revenue³². Approximately 65 percent of data companies have data analysis-related specialists, whereas among general companies that level stands at only 41 percent, according to a 2021 report by the Korea Information Society Development Institute. Among prospective entrepreneurs and startups with poor data processing and analysis capabilities, even if they have data, the direct use to satisfy their needs is limited.

Second, the foundation for the convergence and combination of valuable data is insufficient. Because most of the data platforms focus on data opening and sharing, they are lacking when it comes to providing the analysis and utilization infrastructure needed for actual data utilization. Data platforms focus on basic analysis tools and lack advanced cloud-based analytical functions. Furthermore, although the combination of pseudonymous information is required according to the revision of the Three Acts on Data Policy, the use of pseudonymous information is low given the lack of incentives to combine, the complicated procedures involved, concerns about personal information infringement, and a lack of comprehensive support.

Third, it is difficult to reflect consumer requirements, and the data usability and experience are inadequate. For data to be utilized effectively, they must be developed in line with consumers' intended uses, but the system that can convey and reflect the needs of the consumers is insufficient. It is necessary to expand the base of data service use by allowing data utilization results to be delivered to potential consumers in the public and private sectors.

31 Presentation by Korea Information Society Development Institute (KISDI). (2021)

32 Presentation by IT Industry Promotion Agency (NIPA). (2021)

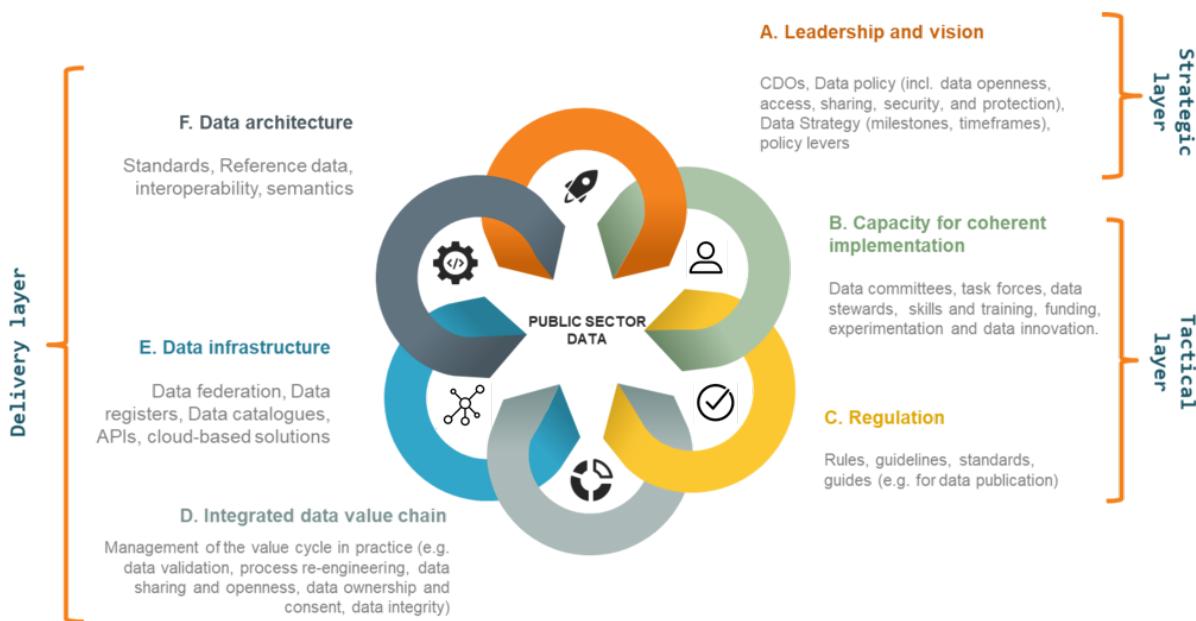
3.4 Future Tasks to Enhance Korea's Data Practices

3.4.1 Establishing a Robust Data Governance Framework

For Korea's data policy to be sustainable and its outcomes improved, the first task would be to establish cohesive and comprehensive data governance. Data governance refers to the vision and goal of data utilization and related policy implementation, participating organizations and division of labor related to data policy, and the structure of the distribution of rights and responsibilities related to the data policy. Good data governance contributes to a common vision and goal among the various stakeholders and institutions and supports cohesive and consistent data-related policy enforcement and coordination. At the same time, the level of control and management to create the value of data is increased by strengthening institutions, regulations, and technological capabilities. It can also increase the level of confidence and create value in the process of data collection, development, storage, protection, processing, sharing, and recycling.

According to the OECD (2019), the components of public sector data governance can be broadly divided into the strategic layer, the tactical layer, and the service delivery layer, with the strategic layer consisting of leadership and vision, the tactical layer consisting of cohesive enforcement capacity of governance and regulation, and the service delivery layer consisting of data value cycle, data infrastructure, and data architecture (Figure 38). Furthermore, the effectiveness and robustness of governance may be realized when the data governance elements are cohesively combined, and each element creates a virtuous cycle.

Figure 38: Public Sector's Data Governance³³



Source: OECD (2019)

Data governance in Korea is decentralized in nature. Let us analyze Korea's data governance in terms of the strategic and tactical layers of public sector data governance suggested by the OECD (2019). The data-related sub-policies and detailed functions such as public data opening, big data platforms, AI training datasets, and personal information protection, which provide the basis of data utilization and data-based administration, not only are scattered across various ministries, committees, and laws, but the status, roles, and functions of related

³³ OECD. (2019). "Digital Government Review of Argentina: Accelerating the Digitalisation of the Public Sector." (<https://doi.org/10.1787/354732cc-en>).

ministries and committees are not systematic and cohesive. The powers and functions of the committees in charge of coordinating functions are ambiguous, and it is difficult for them to enforce requirements (Table 12). Furthermore, because the data policy's process (plan-budget-deliberation-execution-evaluation) is segmented, not only are the data policy plan and budget not linked, but the data policy results, again, cannot be returned to planning, budget, and policy decisions.

Table 12: Status of Governance by Data Policy Category in Korea

		Data-Based Administration	Big Data Platform	AI Data	Personal Information Protection
Ministry with jurisdiction	Ministry of the Interior and Safety	Ministry of Science and ICT	Ministry of Science and ICT	Personal Information Protection Committee (operating under the Office of the Prime Minister)	
Governing laws	Act on Revitalization of Data-Based Administration (December 2020)	Framework Act on Intelligent Informatization (December 2020)	Framework Act on Intelligent Informatization (December 2020) Framework Act on Data Industry Promotion and Utilization (newly established in 2022)	Personal Information Protection Act (September 2011)	
Purpose and key details	Data-based objective and scientific administration	Development of support system for maintaining and strengthening competitiveness across ICT fields including AI, data, and 5G (data production and collection, distribution and use, and standardization and quality improvement, etc.)	Strengthening of the AI competitiveness and the securing of the foundation (development and opening of core data for AI development)	Establishment of personal information protection-related principles, rights of the information subjects, and responsibilities of the state related to personal information protection	
Coordination system	Data Based Administration Activation Committee	Fourth Industrial Revolution Committee (Data Special Committee)	Fourth Industrial Revolution Committee (Data Special Committee)	Personal Information Protection Committee	
	Chairperson: chairperson among the private sector members appointed by the Minister of the Interior and Safety	Operating under the Office of the President (Co-chairpersons: prime minister and civilian members appointed by the president)	Operating under the Office of the President (Co-chairpersons: prime minister and civilian members appointed by the president)	Chairperson: appointed by the president	

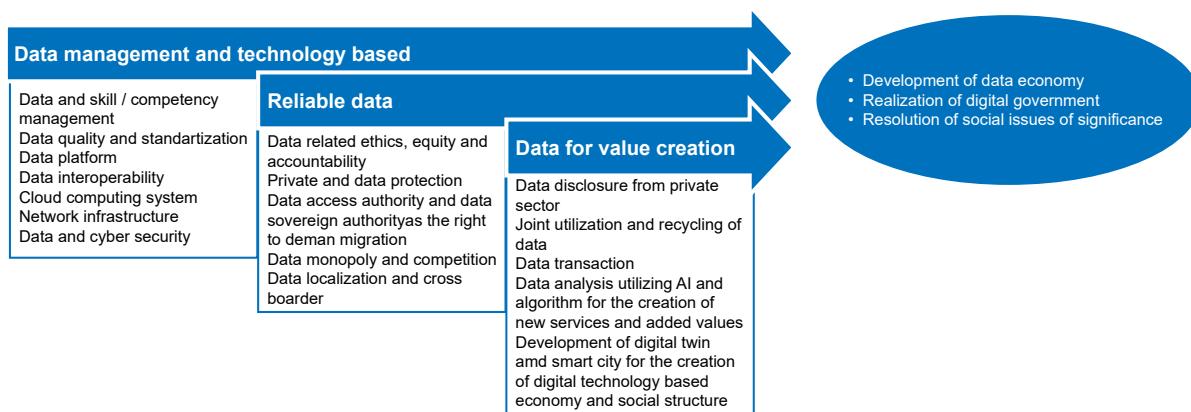
The data-related technology base and management system, including the data value cycle, data infrastructure, and data architecture, which are the service delivery layers, are also fragmented. For example, it is difficult to jointly operate data between the currently operating Open Government Data Platform and related data open platforms. It is difficult to link and use data jointly between the pan-governmental Open Government Data Platform and the data portals built by each local government. Because the e-government systems operated by central administrative agencies, local governments, and public institutions are not connected, it is difficult to jointly utilize data between systems. Furthermore, it is necessary to secure interoperability between such systems and the Open Government Data Platform and big data

platforms because the data analysis system, common data registration system, and meta management system in the government form a separate integrated repository.

In conclusion, for Korea's data policy to be sustainable and achieve the expected effects moving forward, it will be necessary to establish solid data governance that encompasses the public and private sectors. It is also necessary to create a virtuous cycle ecosystem of the entire data cycle based on the establishment of solid data governance, which in turn leads to an increase in the need for integrated data management to improve data utilization performance. Furthermore, it will be possible to improve data policy planning, prevent duplication of data across ministries and the private sector, and efficiently deploy budgets by preparing national data governance.

To this end, a comprehensive approach that considers the data ecosystem is needed (Figure 39). Based on data management and technology infrastructure development, it is also necessary to comprehensively consider data-based administrative services such as MyData, AI-based administrative services, digital twin, and smart city, not to mention data ethics, personal information protection, and data rights. The value-creation effect of data such as national digital transformation must also be considered.

Figure 39: Structure of the Data Ecosystem



Source: Restructured based on the contents of the OECD 2019.

Second, it will be necessary to achieve the integrated management of data, cooperation between data-related ministries and institutions, and the strengthening of coordination systems toward this end, as well as the strengthening of networks with government departments, private companies, and foreign countries through the development of data governance. This rigorous data governance system will help address and coordinate the issues that may arise between stakeholders of the data policy process. It will also enhance the level of using technological infrastructure and links various data platforms across sectors and promote the linkage and integration of different kinds of data, data standardization among other platforms, and knowledge sharing about best practices of platform operation.

3.4.2 Developing a Platform Focused on Consumers and Users

Despite the achievements mentioned above in data disclosure and platform building, there is still criticism that Korea's overall level of data utilization in government and industry is lower than that in advanced countries. Furthermore, it is also argued that the growth rate of data utilization does not reach the level expected at the initial stage of data policy. Among the reasons may be that data disclosure and platform establishment have taken place from a supplier-centric perspective. It has been pointed out that the reason is that core and large-capacity data, which are in high demand as they are opened and focused on individual data

selected by data providers, are opened only to a limited extent. All the data should be opened, based on the principle that data must be open by default.

In data opening and platform development, the consumer- and user-centric principle should be faithfully observed. It is important to collect and open useful and high-quality data. A total service should be provided to improve data accessibility to ensure that it may be easily accessed and used conveniently, and data life cycle activities supported so that collection, processing, and computing resources may be used in the cloud. In terms of the data format and quality, it should be opened in a format that users can easily access, utilize, and analyze. Finally, various data platforms should be linked with the “integrated data map” to ensure that anyone can easily find and use the data to facilitate distribution.

Furthermore, active data-related demand must be initiated. Through the participation of private and public data platform operators, related ministries, institutions, and private sector experts, new demand related to data opening and utilization needs to be created. It is also necessary to consider ways to upgrade the use of vouchers to ensure that consumers, including companies, can freely purchase or receive support for processing in data platforms that meet certain requirements. Moreover, it is necessary to analyze the data usage and purchase status of data platform users and discover customized data by reflecting user opinions. It is also necessary to consider the kind of data brokerage and stewardship activities that can introduce platform data and services to potential users of data.

When considering the subject of data utilization and analysis, rather than thinking with a focus on a small number of decision makers or experts, consideration needs to be given to the general public and public officials in order to create a technology base and culture that promotes their collaboration and discussion with experts. To this end, as presented in Table 13, it may be necessary to differentiate the services for experts from services for the general public and public officials.

Table 13: Public Data Service by User Type

Type	Service	Key Details
Expert	Catalog service	<ul style="list-style-type: none"> Provides detailed information on the data set provided by data platform Provides data using API
	Update notification service	<ul style="list-style-type: none"> Provides notification via text message or email when changes to individual datasets and API services of interest occur
	Advanced search service	<ul style="list-style-type: none"> Provides a variety of detailed search services, such as searching for related words and searching with or without specific words Conducts integrated search for various data sources such as public data opening and information disclosure
Average citizens, average civil servants	Data recommendation service	<ul style="list-style-type: none"> Presents data frequently sought by users and the autocomplete function
	Artificial intelligence search service	<ul style="list-style-type: none"> Recommends analytical techniques and usage data beginners can use according to the purpose of analysis and application domain

3.4.3 Transitioning from a Government-Led to a Market-Friendly Data Policy

In the countries that are leading in the area of data, including the United States, data development and utilization are promoted with a focus on big tech companies, and the data processing market is driving the growth of the data market. On the other hand, in Korea, public data opening, distribution and trading (big data platforms), and practically all aspects of data policy, including the creation and development of new data (AI training datasets), are led and managed by the government. It is undeniable that such government-led data policies

are serving as a catalyst for the utilization of data and vitalization of the ecosystem. However, it is important to review whether such government-led data policies are sustainable over the long term. A market-friendly, virtuous cycle model is needed, in which the data policies can continue to grow and be highly advanced, along with increasing the utilization of platforms and data that have already been developed or are currently in development.

First, in terms of the platform design and operation, it is necessary to strengthen the data platform-based supply system through incentives that encourage the public and private sectors to voluntarily supply and open high-quality data. It is also necessary to focus on strengthening data distribution and transaction capabilities to ensure that users can easily access, conveniently search, and reliably trade with confidence through the data platform.

Second, it is also necessary to build a market-friendly data ecosystem, in which the value of data can be created autonomously within the private sector, while continuously pushing for government support and necessary regulations to secure data that requires a strategic expansion and an increase in consumer convenience and data utilization. To this end, it is necessary to support the building of the social infrastructure, such as by strengthening the data trading environment. In addition, a sub-ecosystem approach should be considered, built according to the characteristics of each industry sector, or facilitating data demand and supply according to the characteristics of each sector.

Third, it is necessary to expand the social foundation for data utilization. To ensure that the supply and demand of data are executed, it will be necessary to develop the data industry and strengthen the data utilization capabilities of the social sector. It will also be necessary to support the start-up efforts of data-related companies and support the use of data-related services by existing companies. It is necessary to provide a data utilization curriculum for students and the general public to expand the base of data development and use, and to link schools and vocational education to ensure they can be connected to relevant jobs and careers.

Finally, detailed plans and roadmaps are needed that define in greater detail the areas and functions in which the government should intervene and the ones in which the private sector should lead and establish an implementation plan based on evaluation criteria such as urgency and marketability. Industry, academia, and research experts in each field should help construct such a roadmap, and based on this, drive changes towards a market-friendly data policy.

IV. A Way Forward for Data Policy Making

4.1 Choosing the Right Time to Deploy the Policy

Korea has emerged as a global leader in leveraging data and AI to drive economic growth and innovation. Over the last several years, the country has made significant progress in establishing and implementing vital enablers such as robust digital data infrastructure.

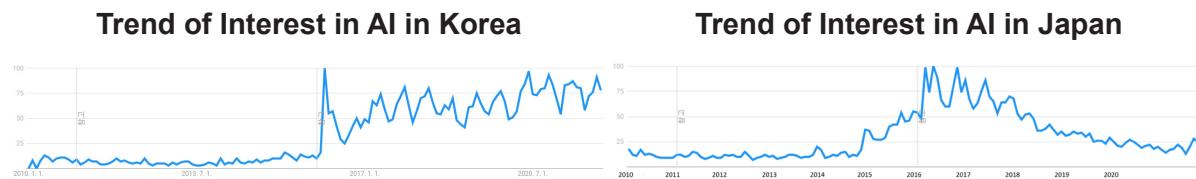
Learning from Korea's experiences could help developing countries in gaining valuable insights into data value realization, developing multiple data platforms, and identifying use cases demonstrating the potential benefits of data-driven AI systems. This knowledge could potentially be used to create successful data governance frameworks, market-friendly data ecosystems, and the social foundations required for data use. Listed below are a few of the lessons acquired.

One of the most important factors in government policy is timing. Regardless of how good a policy is, if it is announced too early or too late it will fade out without having any effect on the market. In this respect, Korea's AI learning data development project could be said to have been announced at a very appropriate time, and thus it succeeded in meeting its goals.

Interest in artificial intelligence in Korea needs to be considered in two periods: before and after the Go tournament between Google's AlphaGo and Lee Sedol of 9-dan rank in 2016. Before this tournament, people were divided on who would win. Lee Sedol was confident enough to say, "It's not about winning, it's about whether it's 5-0 or just giving up a game," whereas Google predicted an even match-up. However, the result of this feud, as we all know, was an overwhelming 4-1 victory by AlphaGo. This sparked great interest in AI among Koreans.

According to Google's trend analysis, illustrated in Figure 40, Koreans' interest in AI was less than 10 on a scale of 100 until 2015, whereas every year since the 2016 match it has risen to a level of 50 or greater. In comparison, a Google trend analysis comparing Korea and Japan shows that even after 2016 the level of interest among the Japanese has been 25 or less, only slightly higher than before the match and significantly lower than among Koreans.

Figure 40: Comparison of Korean and Japanese Interest in Artificial Intelligence, 2010-2021



Source: Google Trends from 2010 to 2021 (<https://trends.google.co.kr/trends/?geo=KR>).

The tournament showed that the Korean government needed a prompt policy not only to maintain public interest in AI, but also to unite the public in order to increase national competitiveness in AI. To this end, in April 2016, the Korean government proposed the Medium and Long-term Comprehensive Measures Implementation Plan for the Intelligent Information Society³⁴ for state affairs review. That September, following the prime minister's decree, the

³⁴ A society where the intelligent information technology, which combines the data created, collected, and accumulated through the advanced information and communication technology infrastructure (IoT, cloud, big data, and mobile) and artificial intelligence, is universally used across all fields of economy, society, and life to create new values.

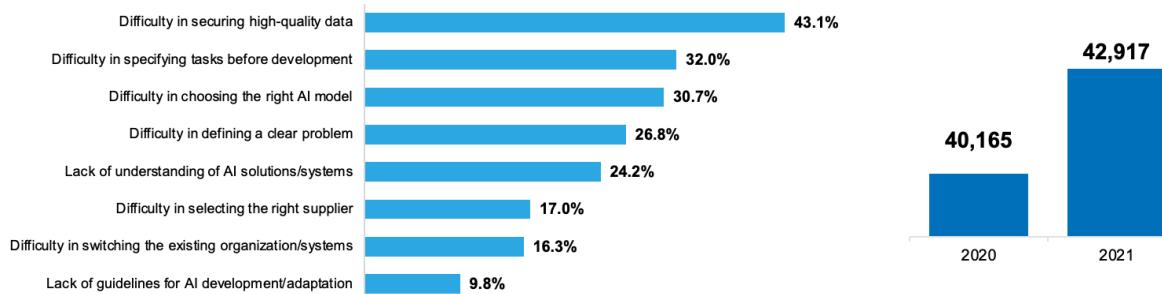
Intelligent Information Society Promotion Unit was launched, consisting of public officials from six ministries (Ministry of Economy and Finance; Ministry of Education; MSIT; Ministry of the Interior and Safety; Ministry of Trade, Industry and Energy; and Ministry of Employment and Labor) and private experts. That December, the Medium and Long-term Comprehensive Measures for the Intelligent Information Society were promptly established and announced. In this measure, the government predicted that intelligent information technology³⁵ would be a growth driver of the Fourth Industrial Revolution and announced the government's mid-to-long-term policy direction to foster intelligent information technology by 2030. Here, the emphasis was placed on an intensive investment in human data resources.

The government began investing in the AI training datasets in earnest based on this confirmed policy direction. But in 2017, the first year of investment, only a small budget of \$2.28 million was allocated. The reason was that the proportion of AI startups in Korea was very low, at 1.6 percent of total startups, and thus even if a large budget was invested to build the AI learning data, the companies that could utilize it were extremely limited. Furthermore, it was important to invest based on an accurate analysis of the market situation because the data that had already been developed would not be helpful or usable unless they were constantly updated to reflect the development of technology.

The scale of the Korean government's investment in developing the AI training datasets has substantially expanded since then, reaching \$235.74 million in 2020. Two factors contributed to such a significant expansion of the budget. The first was the growth of the AI market in Korea. The proportion of AI startups has grown by approximately 1.53 percent each year, reaching 8.8 percent (5.5 times the initial level) in 2020³⁶. As the market grew, the demands of companies in terms of government support also increased, and the common request emerged for the government to build and open AI training datasets through large-scale investments.

Second, the rate of unemployment rose steeply due to the COVID-19 pandemic. According to the Labor Research Institute under the Ministry of Employment and Labor of Korea, during the months of March and April 2020, when COVID-19 broke out, the number of employed people decreased by 1.02 million and the number of workers on temporary leave increased by 990,000, marking a loss of approximately 2 million jobs. The government therefore felt a need to promote a national project that could create large-scale job growth as soon as possible, and the business of building data for AI learning was one that certainly met such conditions.

Figure 41: Factors Hindering Companies' Adoption of AI³⁷ and the Job Creation Effect (persons)



Source: Korea Information Society Development Institute. (2021)

The reason the government decided to build the large-scale AI training datasets by investing \$974.15 million from 2020 to 2022 was that this would create an initial market that could utilize the datasets. It was also decided that if it was delayed further, the technological gap with the advanced AI countries would widen, making it difficult to catch up. Furthermore, it was

³⁵ It is a technology which implements human high level information processing through ICT, where the "intelligence" implemented by artificial intelligence and the "information" based on data and network technologies (IoT, cloud, big data, mobile) are combined.

³⁶ Ministry of SMEs and Startups. (2021). "An Analysis of Changes in the Korean Start-Up Ecosystem."

³⁷ Korea Information Society Development Institute. (2021). "An Analysis of the Factors Hindering the Introduction and Expansion of AI & the Policy Implications."

necessary to promote a large-scale national project that could address the issue of increasing unemployment due to COVID-19 and prepare for the Fourth Industrial Revolution.

4.2 Adopting an Agile and Accurate Decision Making

People often say that “the times make the hero.” The same holds true for good policies. As mentioned earlier, for a policy to be successful, it must be established and implemented at the right time. This requires public officials to make policy decisions. Among the most important pivots in Korea’s ICT development is the correct and prompt policy making by public officials. This has become more important than ever in the recent ICT era, which is changing at the speed of light.

The beginning of a data-driven digital new deal dates back to early 2020. All of Korea’s economic indicators, such as exports, employment rate, and growth rate, fell to their worst levels historically after COVID-19 began affecting the country. The Korean government moved promptly to overcome this national crisis via top-down measures, which it judged to be the most effective approach—due to the fact that bureaucratic government structures slow down the progress of bottom-up policies, whereas top-down decisions can be made close to the speed of private companies, and sometimes even faster.

In January 2020, the Blue House, the final decision making body of the Korean government, instructed the ministries of economy, including the Ministry of Economy and Finance, to establish policies that would help Korea overcome the economic crisis caused by COVID-19.

The Ministry of Economy and Finance carefully reviewed the various policies proposed by each ministry in terms of current and future economic effects. Although the economic effect at the time was the metric used to evaluate how effective the proposed policy was for the real economy, in terms of aspects such as exports, employment rate, and growth rate, it was also important to evaluate whether the proposed policy could contribute to the creation of new industries in the future. The policies selected by the Ministry of Economy and Finance were reported to the Blue House for final decision making, and the Korean version of the New Deal came to life.

The speed of its creation is due to the approximately 300 affiliated organizations that support policy establishment and manage projects under each ministry. These agencies have accumulated deep and wide knowledge of their domains by researching and analyzing issues and trends over a long period of time, which guides their development of appropriate policies.

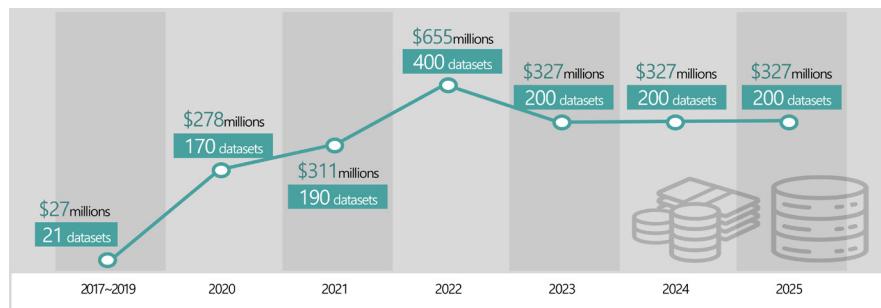
From an interview with Kang Do-Hyeon, Director General in charge of the Ministry of Science and ICT’s Digital New Deal:

“Three to four months were enough to develop the ‘Digital New Deal,’ which is at the core of the ‘Korean version of the New deal.’ The Korean government was able to understand the ICT-related trends over such a short period time, accurately identify market needs, and reflect them in its policies because it has always been ready to establish appropriate policies given the solid collaboration system built and operating by and between the relevant ministries and affiliated organizations for many years, and given that policies centered on the Blue House enable fast decision making practices.”

4.3 Using the Collective Intelligence of “Crowd Workers”

The Korean government announced in August 2022 that it would build 1,300 AI training datasets by investing \$1.97 billion by 2025 based on the Digital New Deal.

Figure 42: The Korean Government's Annual Investment Plan for AI Training Datasets³⁸



Source: National Information Society Agency. (2021)

In line with this plan, the Korean government has invested \$521.23 million over the past two years (2020-2021) to build 340 types of AI training datasets. The original plan of the AI Training Dataset Project established at the beginning of the year was to invest \$30.66 million in 2020 to build 20 types of datasets. However, once the project was selected to be among the core businesses of the Digital New Deal, it was drastically revised to build 150 types of datasets by investing an additional \$229.93 million for approximately five months from October 2020 through March 2021.

The government estimated that 50,000 data labelers were needed to build the additional datasets. However, this represented 7.5 times more individuals than the 7,000 data labelers hired to build the original number of datasets. For this reason, it was impossible to select a business operator through a typical business announcement process and rely on the chosen operator's personnel to complete the labeling tasks. Instead, the government considered introducing a new method in which anyone with digital capabilities could participate in the data labeling work. After gathering various opinions, crowdsourcing was confirmed as the new business method.

Crowdsourcing combines the words "crowd" and "outsourcing," and is a method by which any citizen with digital capabilities can participate in the data collection, refinement, processing, and inspection process whenever desired through an online platform. If the business is promoted by a crowdsourcing method, a company would have the advantage of being able to flexibly hire individuals to carry out the project within a limited period and the crowd worker has the advantage of being able to work at any time, from anywhere. Crowd workers can access the crowdsourcing platform through the internet, whether from home, a cafe, or outdoors, and perform the data labeling at any place work can be performed. One can perform as many tasks as he or she desires given his or her time. Given its advantages, crowdsourcing provides opportunities to work for the socially disadvantaged, such as youth, career-interrupted women, the disabled, and retirees, many of whom are having difficulty finding jobs given the COVID-19 situation and making it possible for them to meet their minimum needs by earning regular income. Therefore, the Korean government was able to gain the trust of citizens and companies even with the COVID-19 situation.

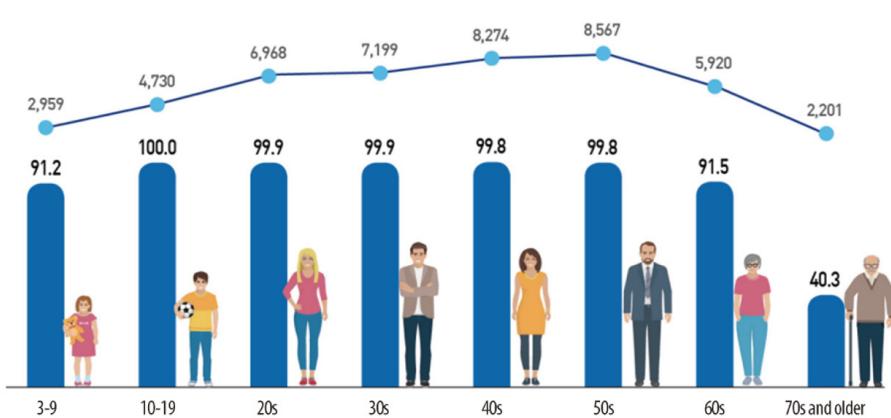
There are a few caveats that must be considered in order to successfully promote a business based on such a crowdsourcing method. First, the national ICT infrastructure must be sufficient. As mentioned, crowdsourcing is a method by which crowd workers access the platform of a crowdsourcing company at any time, from anywhere, and label large files such as images, videos, and audio in real time. Hence, a well-established network must be provided at the national level. Korea ranks first among OECD member countries in terms of high-speed fiber ratio (81.7 percent), mobile broadband usage (24 gigabytes per month), and Internet download speed (156 megabits per second). Thus, any crowd worker who seeks to participate in data labeling work could do so, regardless of their location and available hours.³⁹

³⁸ National Information Society Agency. (2021). "Outcomes of 'The Open Dataset Project."

³⁹ OECD. (2020). "The OECD Digital Economy Outlook 2020."

Second, there must be numerous potential workers with ICT application skills. Unlike AI modeling or programming, data labeling is a task that the general public can easily participate in because the barrier to entry is low. However, one must be able to handle a computer proficiently and be familiar with downloading, installing, and using software. As of July 2020, the proportion of households with internet access in Korea was 99.7 percent, and 91.9 percent of the population aged 3 years and older used the internet for an average of 20.1 hours a week. In particular, the rate of internet use for those in their 20s to 50s, who participated most actively in the data labeling work, turned out to be almost 100 percent, which meant that approximately 31 million people could participate in the data labeling.⁴⁰

Figure 43: Rate of Use of the Internet by Age and Number of Users (%; thousand people; based on population ages 3 and older)



Source: Authors

Third, there must be multiple crowdsourcing companies equipped with platforms that enable a large number of workers to connect and work remotely. It is not easy to produce high-quality AI training datasets by effectively managing tens of thousands of people with different levels of experience and understanding of data labeling through crowdsourcing. Thus, crowdsourcing companies that have accumulated experience and know-how in handling such large-scale AI training datasets are needed. In 2019, when the Korean government decided to promote the large-scale AI Training Dataset Project, 14 crowdsourcing companies with expertise gained through multiple project implementations over many years created an environment in which they could recruit and work with crowd workers through their platforms. Such crowdsourcing companies have increased in number over the past three years as the government has expanded investments in the AI Training Dataset Project, rising by 26.74 percent in 2021. Currently, 92 crowdsourcing companies in Korea can recruit tens to hundreds of thousands of people to perform data labeling work, and as such, have the capacity to fully handle the amount required by the Korean government.

Fourth, there must be a system by which crowdsourcing companies can verify the processing of expenses for crowd workers. The crowdsourcing method is not one in which the business participants are specified from the beginning and only the specified participants perform the process and are paid accordingly. Because of the dynamic nature of the work, and the need to pay expenses accordingly, the potential for the fraudulent handling of expenses is higher than in existing business methods. For example, there may be cases of fraud in which a person who did not perform a task is registered as a task performer to handle the cost or to receive payment in excess of the amount of work done. The government has implemented several expense verification systems to prevent such fraud.

⁴⁰ Ministry of Science and ICT & National Information Society Agency. (2021). "The 2020 Survey on the Use of the Internet."

Table 14: Verification System for Handling Expenses for Crowd Workers

Phase	Detailed contents
Phase 1	Mandatory preparation of standard contract presented by the government for all crowd workers
Phase 2	Full-scale investigation of all work results for the monthly high-payment recipients (more than \$157,000)
Phase 3	Random verification of individual work results via the platform log records
Phase 4	Operation of a website for reporting on fraud involving the handling of crowd worker expenses
Phase 5	Mandatory submission of payment details according to individual workload and full-scale confirmation

4.4 Focusing on Quality, Quality, Quality

Above all, the most important criterion for determining whether the AI learning data project was successful is the quality of the developed data. Recently, most AI models required for the development of AI-based services are not sufficient to resolve all problems but can be fully utilized for major problems by utilizing neural network models disclosed through various documents. Consequently, it may be said that the factor that determines the performance of AI-based services is the quality of the data trained on the AI model.⁴¹ A large amount of data is needed for artificial intelligence, but that in and of itself is not sufficient—AI learned from biased data would cause irreparable damage for researchers and companies, including in the realm of social and ethical issues. Some recent incidents caused by poor data quality have spurred companies to renew their focus on the issue.

Amazon, one of the global ICT companies that is best known to Koreans, was found in 2018 to have a bias in which male applicants consistently scored higher than female applicants during the final simulation process when an in-house AI recruitment program was used, and therefore was discarded. The cause was determined to be Amazon's employee structure. The AI recruitment program was based on the data of employees who had high performance and received good evaluations, and among such developers, males greatly outnumbered females. Because most of the high-performing employees were male, the AI naturally determined that it had no choice but to show preference to male applicants.⁴²

In Korea, the AI chatbot Iruda, created by the startup ScatterLab, caused great social disturbance, and its service was suspended just three weeks after its 2021 launch. Iruda was an AI chatbot set up to resemble a 20-year-old college student. Just as people would communicate with their family, friends, and colleagues through a social media messenger, they would converse with a virtual AI chatbot. Iruda was very popular with teens and those in their 20s, and more than 750,000 people used it in its first two weeks. But as the number of users increased, discrimination against the socially disadvantaged and minorities, such as women, the disabled, and members of the LGBTQ+ community began to surface. Iruda made remarks such as "Lesbians are horrible and creepy," "People with disabilities are inconvenient" and "Seats for pregnant women on the subway are detestable." An analysis found that one of the reasons for the discrimination was the quality of the data. In the process of developing Iruda, ScatterLab learned from the conversations of users on other services it retained without refining this learning. Thus, Iruda learned to mimic those biased online conversations.

There have also been many other cases in recent years in which AI has become compromised by learning low-quality data (including biased data). For example, the chatbot Tay, introduced by Microsoft in 2016, learned from some users to swear and to make sexist, racist, and politically provocative remarks 16 hours after the service was launched. These cases show that the quality of data can also have a tremendous impact on the image and longevity of a company.

⁴¹ Ng, Andrew. (2021). "Forget about building an AI-first business. Start with a mission." MIT Technology Review, April 2021.

⁴² "Why does AI prefer male applicants for Amazon," Hankookilbo. (Oct. 2021). <https://m.hankookilbo.com/News/Read/A2021101409500001667>.

The Korean government has recognized the importance of data quality and has made three types of efforts to improve it since the start of the AI Training Dataset Project. First, four essential quality verification indicators were selected (Table 15) to verify whether the data was properly developed.

Table 15: The Four Essential Quality Verification Indicators

Indicators	Details
Diversity	Checks whether the amount of data is uniform for each category required for AI learning
Syntax accuracy	Checks whether the data structure, format, and input values are entered correctly, and whether any information is missing
Semantic accuracy	Checks whether the data are correctly labeled to identify the object
Validity	Checks whether the performance of the AI model has achieved the presented goal by learning the developed data

Diversity verifies whether data with characteristics like the real-world data satisfy the initially planned learning purpose of AI. For example, if the image data of vehicles on the road are built to develop an AI-based autonomous driving service, various car images that may be seen on the road, such as passenger sedans, trucks, and SUVs, must be collected to enable the autonomous driving service to recognize all potential scenarios and perform autonomous driving missions without errors.

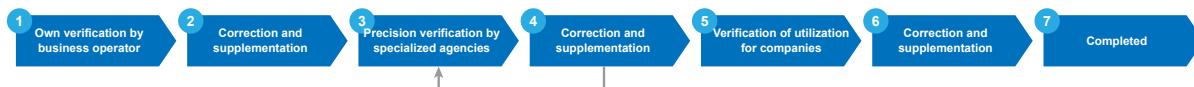
Syntax accuracy also verifies that the data's labeling structure, input value range, format, etc., developed for the AI learning are entered correctly according to the pre-defined syntax rules and that there are no missing data. For example, if there is a dog in the cat folder or if a dog is labeled as a cat it is filtered out in the process of checking for syntax accuracy.

Semantic accuracy verifies whether the labeling for the identification of objects in the source data has been correctly performed. For example, in the process of labeling a cat, if one labels only a part of the cat, such as the head and torso, instead of labeling the entire cat in the data, it will be filtered out in the process of checking for semantic accuracy.

Validity checks whether the performance target is achieved when the developed data are used to train the AI model. For example, if the AI model's performance before data training was 80 percent, and the target after training with the data was set to 90 percent, validity ensures the target performance of 90 percent was satisfied.

Second, in the process of building each data set, a three-phase quality verification procedure is performed. Phase 1 is self-verification. Each business operator should screen the data using the four quality verification indicators before submitting the completed data. The business then must submit the results together with the data. If any of the results are inadequate they must be supplemented and the relevant data submitted. Phase 2 is verification through a data quality verification agency. In this phase, an agency measures the degree of achievement of meeting the goals of the four indicators and supplementary measures are taken for the data which have not reached the goals. Phase 3 is supplementation and re-verification of the data whose quality is found to be insufficient in the previous phase.

Third, before the data is opened, the data is used by AI-specialized institutions and specialized companies and the relevant issues are identified and improved. Regardless of effort, when large-scale data is built there are bound to be errors. Therefore, the companies and institutions that need the AI training datasets attempt to write the data in advance, find specific and practical issues, and improve the quality of data.

Figure 44: Korea's Data Verification Procedures

Source: Authors

Based on such efforts, the Korean government achieved 100 percent in diversity, 99.2 percent in syntax accuracy, 87.2 percent in semantic accuracy, and 81.3 percent in model validity for the 170 types of AI training datasets developed in 2020. This is a level comparable to the data of Microsoft's Common Objects in Context (COCO) or Google's ImageNet, which are recognized as state of the art around the globe.

4.5 Prioritizing Privacy Sacrifice Quality, but Elevate Privacy

Although data quality is a crucial factor in developing AI services, it is superseded by privacy. Korea's Personal Information Protection Act strictly regulates the collection and use of personal data. Collecting or using any personal information or providing it to a third party without the prior consent of the information subject is considered illegal and is subject to sanctions. However, there is a very high risk of violating the act in the process of collecting data to build AI training datasets. For example, in the process of developing the landmark image data that is representative of Korea, such as well-known buildings, public institutions, and tourist attractions, there is a high possibility of collecting the faces of people entering and leaving buildings or passing by tourist attractions. Furthermore, in the process of collecting image data for autonomous driving, it is unavoidable that license plate data owned by individuals will be collected without their consent.

The specialized companies developing AI services want to have data labeled by using data collected without prior consent. In the real world, it is quite natural for people to enter and exit buildings and tourist attractions and for all vehicles on the road to have license plates, therefore more sophisticated AI services can be developed when AI learns such data.

However, the Korean government decided to persuade companies to build AI training datasets within the boundaries of the current laws rather than revising the laws. There were several reasons for this. First, because the social discussion of the protection and use of personal information is still actively underway it would be confusing and impractical to create and promote a separate institutional device to recognize the AI learning data development project as an exception, such as through special laws, presidential decrees, and ministerial notices.

Second, regardless of the thoroughness of the security management system developed by the Korean government to prevent a data leak, the data might still eventually leak through the AI Hub. Experience has taught that once data have leaked online, it is virtually impossible for the government to completely recover them or prevent their circulation. Thus, the Korean government has decided to promote the data development project for AI learning while strictly observing personal information privacy in order to prevent any infringement of privacy through data leaks.

Third, the Korean government has decided to comply with current law to support the global advancement of domestic AI companies. Currently, countries are trying to prevent companies from using unlawfully secured data to develop AI services. Thus, if companies are seeking to launch their own AI services within the next few years, the Korean government may be obliged to prove the legality of the entire life cycle of the data used for the AI learning to the governments of the countries planning to launch them. Accordingly, the government determined that even if the performance of the AI services was degraded through efforts to ensure legal compliance,

over the long term it would be much more beneficial because it removes potential obstacles for domestic companies to enter overseas markets.

Consequently, for all of the AI training datasets developed by the Korean government, the data labeling work was performed targeting data that have undergone a de-identification process to eliminate personal information such as face, name, address, phone number, and license plate for subjects from which prior consent was not secured. For data that may contain personal information, efforts were made to protect the personal information through a complete inspection rather than a sample inspection. As a result, data that may contain personally and nationally sensitive information, such as healthcare and highly precise photos of national land, can be used only in online and offline safe zones, without opening it to the public.

4.6 Building Institutional Capacity for Both Policy Making and Implementation

The role of the NIA, which closely supported the government's policies and projects, was in the background of the Korean government's ability to carry out such a tremendous national project over such a short period of time. The organizational structure of the Korean government changes according to domestic and foreign circumstances every five years when a new administration is inaugurated. For example, the Ministry of Science and ICT was newly established by combining the science and technology and ICT functions in 2013 after removing the Ministry of Information and Communication, which oversaw ICT in 2008.

Changing the organizational structure of the government whenever a new administration is inaugurated would be desirable, in that the government can promptly and intensively respond to the changing external environment. Yet frequent changes in the government's organizational structure would increase the risk of losing the expertise, experience, and know-how that existing ministries have accumulated over many years or decades. If the knowledge, experience, and know-how accumulated by existing ministries are lost in the process of changing the government's organizational structure, the government is likely to lose consistency and continuity in policies and projects resulting in a high possibility of repeating the same mistakes the government made before.

To address such issues, the government has operated an agency to closely support various policies and projects established and promoted by ministries, and to actively utilize them in the process of creating and promoting new businesses. The NIA, operating under the MSIT, has played a large role in promoting the large-scale AI Training Dataset Project focused on the Digital New Deal.

Founded in 1987, the NIA has carried out various national policies related to ICT in connection with the Korean government, business planning, and project promotion, and it has the capability to support the establishment of new policies on ICT-related subjects whenever the government desires.

The NIA has three institutional advantages. The first is the expertise, experience, and know-how accumulated in the process of supporting the government's ICT policy establishment and project implementation over the past 36 years that is embedded into its organization. Given the experience gained from initiatives ranging from the practical management of the National Basic Information System Project, to supporting the planning of the recent Digital New Deal and promoting the data-related projects at its core, it would be no exaggeration to say that all of Korea's ICT-related expertise, experience, and know-how are incorporated within the NIA.

The second advantage is the vast human network retained by the NIA. The NIA is building a strong human network with the government, businesses, and academia alike. A good policy is made when the best experts come together, but creating the best policies is only possible with an organization that can also coordinate their opinions. The NIA always communicates

with Korea's leading experts in each field, such as medical care, education, environment, and national defense, while discussing and analyzing the latest technologies and services and building knowledge together. This allows the NIA to establish and implement various ICT policies required by the government promptly and effectively.

Third, because the NIA has not only supported policy development but also has carried out major projects for implementation, it is possible for the agency to support the development of realistic and specific policies in consideration of market conditions. Through the promotion of various projects, the NIA was able to learn about market status with precision (issues, obstacles, and alternatives), and by incorporating such experiences into the policy making process was able to support the development of realistic and specific policies.

In summary, by examining Korea's data practices, developing countries can gain valuable insights and guidance to shape their own data policies and strategies. This will enable them to harness the transformative potential of data and AI for sustainable development and economic growth.

Bibliography

1. Address by President Moon Jae-in at Opening of 21st National Assembly (July 16, 2020). (<http://webarchives.pa.go.kr/19th/english.president.go.kr/BriefingSpeeches/Others/851>).
2. Dong-hyun Lee, & Heo Jeong. (2018). "Labor Market Forecast of Promising Software Areas." Issue Report, Volume 1, Software Policy & Research Institute.
3. Cole McFaul et al. (2023). "Assessing South Korea's AI Ecosystem". CSET Data Brief: August 2023. Center for Security and Emerging Technology.
4. Evan A. Feigenbaum and Michael R. Nelson. (2021). "The Korean Way with Data: How the World's Most Wired Country Is Forging a Third Way." Carnegie Endowment for International Peace
5. Evan A. Feigenbaum and Michael R. Nelson. (2022). "Data Governance, Asian Alternatives: How India and Korea are Creating New Models and Policies." Carnegie Endowment for International Peace
6. Government of the Republic of Korea. (2017). "Mid and Long-term Comprehensive Measures for Intelligent Information Society."
7. Government of the Republic of Korea. (2018). "Data Industry Revitalization Strategy."
8. Government of the Republic of Korea. (2018). "Intensive Training Plan for Leading Human Resources in the 4th Industrial Revolution."
9. Government of the Republic of Korea. (2019). "National Strategy for Artificial Intelligence."
10. Government of the Republic of Korea. (2020). "Korean New Deal."
11. "Government builds 10 big data platforms and 100 centers... 151.6 billion won invested over 3 years." UPI News. (2019). (<https://m.upinews.kr/newsView/1065617120735279>).
12. Hyun-jin Lee, & Mi-hye Lee. (2021). "Current Status of Artificial Intelligence Industry and Policies to Promote Major Countries." K New Deal Industry INSIGHT Report, Export-Import Bank of Korea.
13. Institute of Information & Communications Technology Planning & Evaluation & Electronics and Telecommunications Research Institute. (2021). "Analysis of National ICT Innovation Capabilities in 2020."
14. Jacob Arturo Rivera Perez, Cecilia Emilsson and Barbara Ubaldi. (2020). "OECD Open, Useful and Re-usable data (OURdata) Index." OECD Policy Papers on Public Governance No.1, March 2020.
15. Jieun Oh, H. J. Lim, H. J. Lee, & S. S. Shin. (2021). "A Policy Study on the Utilization of Public Data as a Crisis Management Tool for COVID-19." Journal of Science and Technology Policy, Volume 4, 43-69.
16. Jungsoo Park, Soonae Park, Jung-Ju Lee, and Sun-Ha Kim. (2003). "Review of the Adequacy of the Informatization Budget and Efficient Resources Allocation." University of Seoul Law and Administration Research Institute.
17. Korea Information Society Development Institutue. (2021). "An Analysis of the Factors Hindering the Introduction and Expansion of AI & the Policy Implications."
18. Korea Data Agency. (2022). "2021 Data Industry Survey." March 31, 2022.
19. Korea Data Agency. (2022). "Big Data Academy." (<https://dataonair.or.kr/bigdata>).

20. Korea Data Agency. (2023). "About MyData."
21. Korea Labor Institute. (2020). "Evaluation of the Labor Market and 2021 Forecast." Labor Review, December Issue.
22. Kyung Sin Park and Natalie Pang. (2022). "Data Innovations and Challenges in South Korea: From Legislative Innovations for Big Data to Battling COVID-19." Konrad Adenauer Stiftung.
23. Ministry of Science and ICT. (2018). "R&D Strategy for Artificial Intelligence."
24. Ministry of Science and ICT. (2015). "Software-driven University Initiative."
25. Ministry of Science and ICT. "The 2018 Status Survey on Data Industry."
26. Ministry of Science and ICT & Korea Data Agency. (2021). "2020 A Survey on Data Industry Status."
27. Ministry of Science and ICT & National Information Society Agency. (2021). "A Survey on Internet Usage."
28. Ministry of Science and ICT & National Information Society Agency. (2021). "The Report on the Digital Divide."
29. Ministry of Science and ICT and the National Information Society Agency. (2022). "The Open Dataset Project"
30. Ministry of SMEs and Startups. (2021). "An Analysis of Changes in the Entrepreneurship Ecosystem in Korea."
31. Ng, Andrew. (2021). "Forget about building an AI-first business. Start with a mission." MIT Technology Review, April 2021.
32. National Computerization Agency. (1997). "A Comprehensive Evaluation of National Basic Information System Project in Stage 2."
33. National Information Society Agency (NIA). (2021). "Outcomes of 'The Open Dataset Project."
34. National Information Society Agency (NIA). Statistics on Opening the Public Data
35. National Information Society Agency (NIA). Statistics on the Use of AI Hub and Internal Data and materials
36. Open Data Strategy Council. (2022). The Public Data Strategy Committee. (<https://www.odsc.go.kr>).
37. Open Data Strategy Council. (2023). Introduction on Open Data. (<https://www.odsc.go.kr/eng>).
38. OECD (2019). "Digital Government Review of Argentina: Accelerating the Digitalisation of the Public Sector." (<https://doi.org/10.1787/354732cc-en>).
39. OECD. (2020). "The OECD Digital Economy Outlook 2020."
40. OECD. (2022). "Towards an Integrated Health Information System in Korea."
41. Open Government Partnership (2023). "About Open Government Partnership." (<https://www.opengovpartnership.org/about/>).
42. Presentation by Korea Information Society Development Institute (KISDI). (2021).
43. Presentation by IT Industry Promotion Agency (NIPA). (2021).
44. "Reviewing the 'identification number' system for public-private led big data platforms." Zdnet. (2021). (<https://zdnet.co.kr/view/?no=20210611163147>).
45. Soh, Hoon Sahib; Koh, Youngsun; Aridi, Anwar. (2023). "Innovative Korea: Leveraging Innovation and Technology for Development." World Bank. Washington, DC

46. Sungsoo Hwang and Joon Mo Abn. (2022). "Digital Government and Public Data Act of Korea."
47. Seok-jin Eom. (2022). "Legal and Institutional Arrangements for Digital Government". 2022 KIPA Module Series. (https://www.kipa.re.kr/synap/skin/doc.html?fn=FILE_0000000000162450&rs=/convert/result/201512/)
48. The National Assembly Research Service. (2021). "Issues and Challenges of Using Artificial Intelligence through 'iLUDA.' Issue and Point, Volume 1779.
49. "Why does AI prefer male applicants for Amazon," Hankookilbo. (Oct. 2021). (<https://m.hankookilbo.com/News/Read/A2021101409500001667>).
50. World Bank. (2021). "Harnessing Artificial Intelligence for Development on the Post-COVID-19 Era: A Review of National AI Strategies and Policies." Analytical Insights. World Bank, Washington DC.
51. World Bank (2022). "Harnessing Trustworthy Artificial Intelligence in Korea: Reshaping the Digital Future in a Post-COVID era." Webinar on January 26, 2022. (<https://www.worldbank.org/en/events/2022/01/25/harnessing-trustworthy-ai-in-korea-to-reshape-the-digital-future-in-a-post-covid-era#1>)
52. Yong-Gwan Jeong & Yoo-Jung Kim. (2004). "An Exploratory Study on Analyzing the Multi-Dimensional Effectiveness of Broadband Network of Korea." National Computerization Agency, Information Systems Review, Vol.6, No.2, December 2004.

