Seok-Jin Eom · Yoon-seok Ko ·
Jeong Eun Park · Keon Chul Park

Enabling Data-Driven Innovation and AI Government in Korea

Springer

Seok-Jin Eom, Yoon-seok Ko, Jeong Eun Park and Keon Chul Park

Enabling Data-Driven Innovation and AI Government in Korea



Seok-Jin Eom
Graduate School of Public Administration, Seoul National University, Seoul, Korea (Republic of)

Yoon-seok Ko
National Information Society Agency, Daegu, Korea (Republic of)

Jeong Eun Park
National Information Society Agency, Daegu, Korea (Republic of)

Keon Chul Park
Department of Business Administration, Tech University of Korea, Siheung-si, Korea (Republic of)

ISBN 978-981-95-9174-9 e-ISBN 978-981-95-9175-6
https://doi.org/10.1007/978-981-95-9175-6

Translation from the Korean language edition: “Hanguk-ui deiteo jeongchaek-gwa AI jeongbu” by Seok-Jin Eom et al., © HanulMPlus Inc. 2025. Published by HanulMPlus Inc.. All Rights Reserved.

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2026

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Acknowledgement

The first part of this book is based on the original data of the World Bank Group, and some of its contents have been summarized, edited, and reconstructed. In addition, it makes it clear that the content of this work does not reflect the official views of the World Bank or its executives and employees, nor does it imply that the World Bank supports or warrants this work, author, or related institution.

Introduction

The government is not only the greatest patron of information and communication technology (ICT) but also its largest consumer (Fountain, 2001; Mazzucato, 2015). Since the 1990s, as the Internet became commercialized and widely available, the integration of ICT into government operations has led to the rise of the concept of "e-government." This was not only a strategy for managing internal human and material resources but also a key means of improving public service quality for citizens and businesses, transforming the political relationship between the government and the people.

With the introduction of new technologies such as mobile and social media (Web 2.0), the way governments operate and deliver public services continued to evolve, along with the concept of government itself. New forms emerged, such as the "mobile government" based on wireless Internet and smartphones, "social media government" leveraging two-way communication, and "open government" promoting proactive disclosure of government information and data. Initiatives such as the Obama administration's "We the People" petition platform exemplify the concept of "Government as a Platform" (O'Reilly, 2012), aiming to move beyond the limitations of traditional e-government by emphasizing the use of digital platforms to enhance government operations, coordinate functions rationally, promote inter-agency collaboration, and innovate public service delivery and citizen-government relations (Eom et al., 2020).

Today, new technologies such as artificial intelligence (AI), cloud computing, big data, and virtual reality (VR) are redefining the future of government and administration. These technologies represent not just a continuation of past ICT developments but a paradigm shift (Eom et al., 2020; Schwab, 2016). Earlier ICT advancements aimed to increase efficiency through process optimization. Now, however, these emerging technologies demonstrate cognitive and computational capabilities that surpass human abilities. The expansion of the Internet has moved beyond connecting people to also linking people with things and things

with other things. This convergence of cognitive capacity, connectivity, virtual reality, and computational power has given rise to autonomous vehicles, drones, and robots that are now widely deployed. These technological advances and their societal impact bring both optimistic visions of the future and concerns about the collapse of existing norms and systems, prompting extensive debate (Eom et al., 2020; Susskind, 2020).

The Korean government has not lagged behind in these global trends. Since the introduction of Korea's first computer (IBM 1401) into the Economic Planning Board's Statistics Bureau on June 24, 1967, Korea has implemented a series of digital government initiatives: the Administrative Computerization Project in the 1970s, the National Backbone Computer Network Project in the 1980s, the High-Speed National Information and Communications Network and Informatization Promotion Projects in the 1990s, and the "11 e-Government Projects" and "31 e-Government Roadmap Projects" in the 2000s. These have been accompanied by investments in education for public officials and citizens, the development of information infrastructure, and administrative process reengineering. Accordingly, substantial tangible and intangible efforts have been made.

Since the 2010s, as the transition to a data-driven economy gained momentum, the Korean government implemented various policies and projects to drive digital transformation in the economy, society, government organizations, and administrative processes. These include building a national data ecosystem, developing diverse data platforms, revising legal frameworks, and initiating policies to incorporate emerging technologies such as machine learning, large language models (LLMs), and generative AI into government.

Despite the global acceleration of technological advancement, a multi-dimensional analysis of how the Korean government has adopted ICT and transformed administrative practices is lacking. In particular, few studies have comprehensively analyzed Korea's post-2010 data policies, including their achievements and limitations, through lenses such as data governance, legal frameworks, technology infrastructure,

and data itself. Moreover, the implementation and utilization of AI in government and administration are no longer future goals but current imperatives. Therefore, it is now crucial to discuss the necessary institutional and technical strategies, as well as the multi-dimensional components and integrations required for implementing AI government. These challenges are specific and require a combination of policy and technology, yet the academic discourse in Korea remains largely general and abstract.

This book is written with the following objectives in mind. First, focusing on Korea's data policies since the 2010s, Part 1 outlines the goals of these policies aimed at transitioning to a data economy, along with their governance structures, legal and institutional frameworks, technical infrastructure, platforms, and sector-specific challenges (Chapter 1). Chapter 2 analyzes three core national data platforms: (1) the Public Open Data Portal, (2) Big Data Platforms, and (3) the AI Hub. Chapter 3 evaluates the outcomes and limitations of Korea's data policies and proposes future directions. Chapter 4 presents policy insights derived from the implementation of these policies, offering practical guidance for developing countries pursuing similar initiatives. Through this analysis, the book demonstrates that Korea's data policy has evolved with distinct characteristics compared to those of leading countries.

Second, this book defines "AI government" as a government that adopts and utilizes AI and related technologies in administration and public operations, and proposes a vision and implementation strategies for such a government (Part 2). Specifically, it examines the characteristics AI-driven public services should possess and explores the overall administrative model an AI government should pursue (Chapter 5). Moreover, to ensure that administrative services and internal management within an AI government function effectively, the book discusses the required technological infrastructure, data, and data-learning processes through the AI Government Foundation Model (Chapter 6) and AI Government Implementation Strategies (Chapter 7).

Finally, the technological and policy-related challenges and tasks involved in implementing an AI government are addressed in Chapter 8.

We hope this book, which examines Korea's data policy and the implementation of AI government, contributes to the meaningful digital transformation of the Korean government and public administration. We also expect it to provide a basis for reflection and discussion on Korea's data policy achievements to date. Finally, we hope that Korea's experience with data policy and AI government extends beyond being a uniquely Korean case and contributes to global discourse and shared understanding.

Contents

Part I Korea's Data Policy

1 Overview of Data Policy

1.1 Historical Background and Development of Data Policy

1.1.1 National Informatization and Construction of the National Backbone Computer Network

1.1.2 Transition to E-Government and the Data Economy

1.2 Key Elements of Data Policy

1.2.1 Development of Data Platforms and Large-Scale Datasets

1.2.2 Strengthening Data Policy Governance and Institutional Foundations

1.2.3 Data Talent Development Policy

2 The Three Major Data Platforms

2.1 Overview

2.2 Open Government Data Portal

2.2.1 Overview

2.2.2 Key Services Provided by the Public Data Portal

2.2.3 Technological Foundation of the Open Government Data Portal

2.2.4 Data Registration and Provision Methods

2.2.5 Metadata Management and Data Integration

2.2.6 Development of the Next-Generation Open Government Data Portal

2.3 Big Data Platform

2.3.1 Overview

2.3.2 Key Services Provided by the Big Data Platform
2.3.3 Technological Foundation of the Big Data Platform
2.4 AI Hub
2.4.1 Overview
2.4.2 Core Services Provided by AI Hub
2.4.3 Technological Foundation of AI Hub
3 Achievements and Challenges of Korea's Data Policy
3.1 Achievements of Data Policy
3.1.1 Achievements of Open Government Data
3.1.2 Achievements of Big Data Platforms
3.1.3 Achievements of AI Hub Development
3.2 Characteristics and Success Factors of Korea's Data Policy
3.2.1 Characteristics of Data Policy
3.2.2 Success Factors of Korea's Data Policy
3.3 Challenges and Limitations of Korea's Data Policy
3.3.1 Issues with Data Collection and Quality
3.3.2 Data Distribution and Transactions
3.3.3 Lack of Private-Sector Utilization and Innovation
3.4 Future Challenges of Data Policy
3.4.1 Establishing Robust Data Governance
3.4.2 Building Demand- and User-Centered Platforms
3.4.3 Transition from Government-Led to Market-Friendly Data Policy
4 Lessons Learned from Korea's Data Policy Implementation
4.1 Timeliness of Policy Implementation

4.2 Swift and Correct Decision-Making and Policy Implementation

4.3 Utilization of Collective Intelligence: “Crowd Workers”

4.4 Data Governance and Quality Management

4.4.1 Establishing Dedicated Organizations for Policy Planning and Execution

4.4.2 Strategy for Improving Data Quality: Quality, Quality, and Quality!!!

4.4.3 Strengthening Personal Data Protection and Ethics (Sacrifice Quality but Elevate Privacy)

Part II AI Government

5 Characteristics of AI Services and the AI Government

5.1 Concept, Working Mechanism, and Components of AI

5.1.1 The Concept and Working Mechanism of AI

5.1.2 Core Components of AI: Key Elements of AI Operation

5.2 Characteristics of AI Services

5.2.1 Overview

5.2.2 Personalized Services

5.2.3 Specialized Services

5.2.4 Self-evolving Services

5.3 Characteristics of Administrative Functions in an AI Government

5.3.1 Precision Administration

5.3.2 Augmented Administration

5.3.3 Connective Administration

6 Implementation of the Government AI Foundation Model as the Brain of the AI Government

6.1 Overview

6.2 Selection of Reference Model

6.3 Collection of High-Quality Data

6.4 Data Cleaning and Processing

6.5 Model Optimization

6.6 Pre-training

6.7 Fine-Tuning

7 Implementation Strategies for AI Government Agents

7.1 Overview

7.2 Single Access Point for AI Government: The AI Government Agent

7.3 The Brain of the AI Government: Government AI Platform (GAIP)

7.4 Search Engine for All Government Services: Whole-of-Governmental Service Catalog

7.5 Small, Lightweight, and Modular LEGO-Like Government Services

7.6 High-Quality Machine-Readable Data Based on a Whole-of-Government Classification System

7.7 Problem Solving Through Modular Combination of LEGO-Like AI Government Services

8 Challenges and Tasks for Implementing the AI Government

8.1 Overview

8.2 Verification of Government AI Models

8.3 Transition to Machine-Readable Data

8.4 Establishing a National-Level Data Classification Model

8.5 National Data Quality Verification for Ensuring Data Integrity

8.6 Transition to a LEGO-Like Modular System

8.7 Validation of Virtual Service Appropriateness

References

Index

List of Figures

Fig. 1.1 Changes in the Korean Government's ICT policy

Fig. 1.2 Korea's data ecosystem. *Source* Compiled by the author

Fig. 1.3 Korea's Digital New Deal policy: The Data Dam. *Source*: Compiled by the author

Fig. 1.4 Shortage rate of big data professionals by job category (2018–2023). *Source* Ministry of Science and ICT (2019)

Fig. 1.5 Korea's data talent development system. Source: Presidential Committee on the Fourth Industrial Revolution (2018).

Fig. 2.1 Conceptual diagram of Open Government Data Portal. *Source* National Information Society Agency (NIA) (2019a)

Fig. 2.2 Conceptual diagram of Open Government Data Portal. *Source* National Information Society Agency (NIA), Republic of Korea (2019b)

Fig. 2.3 National Data Map. *Source* Open Government Data Portal website

Fig. 2.4 Comparison of format levels for machine-readable data. *Source* Ministry of the Interior and Safety (2021)

Fig. 2.5 Open Government Data Openness Framework. *Source* Ministry of the Interior and Safety (2021)

Fig. 2.6 Open API integration strategy based on multi-cloud (public and private) infrastructure. *Source* National Information Society Agency (NIA), Republic of Korea (2024b)

Fig. 2.7 Structure and operational framework of the Big Data Platform. *Source* Jeong, Junhwa (2020)

Fig. 2.8 Usage permissions by data type in the Big Data Platform. *Source* Big Data Platform Website (www.bigdata-cancer.kr)

Fig. 2.9 “City in Crisis” screen of the Coastal Big Data Platform’s Safety Service. *Source* Coastal Big Data Platform Website (www.bigdata-coast.kr)

Fig. 2.10 Integrated Data Map homepage. *Source* Integrated Data Map homepage (www.bigdata-map.kr)

Fig. 2.11 Overview of Big Data Platform-based analytics support program. *Source* National Information Society Agency (NIA), Republic of Korea (2024d)

Fig. 2.12 Big data sharing and integration framework. *Source* Statistics Korea, Statistical Education Center (2019)

Fig. 2.13 Lifecycle-based quality management framework for AI training dataset project. *Source* National Information Society Agency (NIA), Republic of Korea (2023b)

Fig. 2.14 AI Hub homepage interface. *Source* AI Hub Website (www.aihub.or.kr)

Fig. 2.15 AI computing service management framework of AI Hub. *Source* AI Hub Website (www.aihub.or.kr)

Fig. 2.16 Overview of AI Hub's Safe Zone. *Source* AI Hub (n.d.)

Fig. 3.1 Trends in the provision of open government data. *Source* Statistics Korea; e-Nara Index

Fig. 3.2 Trends in the utilization of open government data. *Source* Statistics Korea; e-Nara Index

Fig. 3.3 Trend in the proportion of public data in open format. *Source* National Information Society Agency (NIA) (2022); Statistics on the Status of Open Government Data

Fig. 3.4 2023 OECD Open Government Data rankings. *Source* Ministry of the Interior and Safety (2024)

Fig. 3.5 Number of visitors and registered users of AI Hub (2018–2020). *Source* National Information Society Agency (NIA) (2021); Statistics on the Use of AI Hub and Internal Data

Fig. 3.6 AI Hub key service usage statistics (2018–2020). *Source* National Information Society Agency (NIA) (2021); AI Hub and Internal Data Usage Statistics

Fig. 3.7 Data governance for the public sector. *Source* OECD (2019)

Fig. 3.8 Data ecosystem composition. *Source* Reconstructed based on OECD (2019)

Fig. 4.1 Barriers to AI adoption by enterprises and its effects on job creation. *Source* Lee and Kim (2021)

Fig. 4.2 The Government's annual investment plan for AI training datasets. *Source* Author's own work

Fig. 4.3 Korea's data verification procedure. Source: Ministry of Science and ICT & NIA (2023b)

Fig. 5.1 The history of AI development. *Source* Author's compilation

Fig. 5.2 Working mechanism of AI. *Source* Author's compilation

Fig. 5.3 Dynamic inference. *Source* Google Developers (n.d.)

Fig. 5.4 Static inference. *Source* Google Developers (n.d.)

Fig. 5.5 Traditional AI versus machine learning-based AI. *Source* Lee (2016)

Fig. 5.6 Key components by stage of AI operation. *Source* Compiled by the author

Fig. 5.7 Increase and changing characteristics of data. *Source* National Information Society Agency (NIA) (2011)

Fig. 5.8 Structure of an artificial neural network. *Source* Lee (2016)

Fig. 5.9 Fraud detection case using XAI (feat. mini card). *Source* Electronic Times (2024)

Fig. 6.1 Development process of government Foundation Models for AI. *Source* Reconstructed from Bi et al. (2024)

Fig. 6.2 Major labeling methods for unstructured data. *Source* AIHub (www.aihub.or.kr)

Fig. 7.1 AI government service framework. *Source* Author's compilation

Fig. 7.2 Analog-style library book search desk. *Source* https://www.pexels.com/ko-kr/photo/6333743/

Fig. 7.3 Stack remote monitoring system. *Source* Korea Environment Corporation (www.keco.or.kr)

Fig. 7.4 Conceptual diagram of interdevice connectivity for the AI government agent. *Source* Author's compilation

Fig. 7.5 Conceptual diagram of service connectivity of the AI government agent. *Source* Author's compilation

Fig. 7.6 Screenshot of Government24. *Source* Government24 (www.gov.kr)

Fig. 7.7 Procedure for creating Government AI models. *Source* Compiled by the author

Fig. 7.8 Service classification by government function (example). *Source* Ministry of the Interior and Safety (2024)

Fig. 7.9 Time taken for online services to reach one million users. *Source* Statista (2023)

Fig. 7.10 Comparison between waterfall-based e-government and module-based AI government. *Source* Ministry of the Interior and Safety [MOIS] & Korea Local Information Research & Development Institute [KLID], 2024).

Fig. 7.11 Private cloud environment of e-government vs. hybrid cloud environment of AI government. *Source* Sitesbay (n.d.) and Lee (2004)

Fig. 7.12 Example of government function/organization-based data and metadata classification system. *Source* Ministry of the Interior and Safety (2024)

Fig. 7.13 Virtual service created through the combination of modular (legoLEGO-block style) AI government services. *Source* Author's compilation

Fig. 8.1 Example of a metadata classification system using DCAT. *Source* Author's compilation

Fig. 8.2 Concept and example of function-based data classification system. *Source* Author's compilation

Fig. 8.3 Data quality verification methods. *Source* Author's compilation

List of Tables

Table 1.1 Overview of the big data master plan and big data service activation strategy

Table 1.2 Overview of the Internet New Industry Promotion Plan

Table 1.3 Overview of the Data Industry Activation Strategy

Table 1.4 Key contents of the Data Industry Promotion Act

Table 1.5 National Data Policy's implementation system

Table 2.1 Key provisions of the public data act

Table 2.2 Key implementation framework of public data policy

Table 2.3 Structure and data products of the 21 sectoral Big Data Platforms

Table 3.1 Performance of national core data openings (Phases 1–4)

Table 3.2 Public data services by user type

Table 4.1 Verification system for cost processing of crowd workers

Table 4.2 Four essential quality verification indicators for AI training data

Table 5.1 Data processing procedure

Table 5.2 AI learning methods

Table 5.3 Examples of explainable AI applications by sector

Table 5.4 Examples of global cloud services for AI

Table 5.5 Differences between artificial general intelligence (AGI) and narrow artificial intelligence (Narrow AI)

Table 5.6 Stages of general AI development and examples of major services

Table 5.7 Key functions of an AI assistant (examples)

Table 5.8 Examples of commercialized AI-based professional services

Table 6.1 Representative foundation AI models

Table 6.2 Representative open-source large language models (LLMs)

Table 6.3 Representative methods for data deduplication

Table 6.4 Comparison of major labeling methods for unstructured data

Table 6.5 Key evaluation metrics for labeled data

Table 6.6 Key hyperparameters

Table 7.1 Rule-based system vs. AI-based system

Table 7.2 Government AI-based services vs. Government AI Platform

Table 7.3 Datasets used to train the base model of ChatGPT

Table 7.4 Data requirements for creating Government AI fine-tuning models

Table 7.5 Components of a Whole-of-Government Service Catalog (example)

Table 8.1 Quality measurement indicators for data used in training government AI models

Table 8.2 Detailed indicators by data quality measurement category (examples)

About the Authors

Seok-Jin Eom is a professor of Graduate School of Public Administration at Seoul National University, Korea. He received his Ph.D in public administration from Seoul National University. He has published a number of papers in academic journals including *Government Information Quarterly*, *Public Management Review*, *Administration and Society*, and *Perspectives on Public Management and Governance*. His current research interests include evolution of e-governance in the digital era, the digital transformation of government as well as the intellectual history of public administration research and education in Korea.

Yoon-seok Ko is a vice president at National Information society Agency (NIA), and he has been working for NIA from 2003. He majored in Computer Science for undergraduate at the University of Georgia in the U.S. and got M.S. in Information Systems at Yonsei University in Korea. His specialty is e-government, ICT convergence, and data. He has planned and managed nationwide e-government projects such as Government Business Management System Project and Government Information Data Center Project. He has also worked at Presidential Committee on Government Innovation and Decentralization (PCGID) to oversee presidential agenda, especially e-government projects. He is currently serving as a member of the Presidential Digital Platform Government Committee Task Force and as a senior advisor at the World Bank.

Jeong Eun Park is a researcher fellow at National Information society Agency (NIA). She received her Ph.D in public administration from Ewha Womans University. She served as a vice president of the department of ICT policy at NIA. She also served as an ICT advisor for

Gyeonggi Province. She was a visiting professor at Kyung Hee University and vice president of the Internet Ethics Association.

Keon Chul Park is an assistant professor in the Department of Business Administration at Tech University of Korea, appointed in September 2025. He currently serves as the director of the university's Institutional Research (IR) Center and leads the Urban AI Engineering House (EH), where he focuses on developing AI applications for urban environments. Prior to his current academic appointment, Dr. Park served as the head researcher of the AI X Research Center at the Advanced Institute of Convergence Technology, Korea. He received his M.S. and Ph.D. degrees in Information Systems from Yonsei University. His research interests center on urban computing and urban intelligence, with a focus on applying advanced data technologies, such as big data and AI, to metropolitan areas. His current research emphasizes enhancing urban services and infrastructure through machine/deep learning. Dr. Park is dedicated to leveraging technology to solve complex urban challenges and improve the quality of life.

Part I
Korea's Data Policy

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2026
S.-J. Eom et al., *Enabling Data-Driven Innovation and AI Government in Korea*
https://doi.org/10.1007/978-981-95-9175-6_1

1. Overview of Data Policy

Seok-Jin Eom[1] ✉, Yoon-seok Ko[2] ✉, Jeong Eun Park[3] ✉ and Keon Chul Park[4] ✉

(1) Graduate School of Public Administration, Seoul National University, Seoul, Korea (Republic of)
(2) National Information Society Agency, Daegu, Korea (Republic of)
(3) National Information Society Agency, Daegu, Korea (Republic of)
(4) Department of Business Administration, Tech University of Korea, Siheung-si, Korea (Republic of)

✉ **Seok-Jin Eom (Corresponding author)**
Email: sjum21@snu.ac.kr

✉ **Yoon-seok Ko**
Email: ysko@nia.or.kr

✉ **Jeong Eun Park**
Email: jpark0070@gmail.com

✉ **Keon Chul Park**
Email: parkkc07@tukorea.ac.kr

Keywords Korea data policy – Data economy – Digital new deal – Data governance – Information infrastructure – Data talent development

1.1 Historical Background and Development of Data Policy

The Korean government constructed a backbone of computer networks by securing infrastructure through national informatization. It then laid the foundation for ICT by building e-government systems and establishing a high-speed national communications network. Subsequently, the government promoted a transition toward a data-driven economy by fostering emerging technologies such as big data, the Internet of Things (IoT), cloud computing, and AI. The aim was to enhance administrative transparency while revitalizing the data economy through proactive government-led data openness and utilization policies.

As illustrated in Fig. 1.1, this process is the result of the organic interaction among core components such as infrastructure, policy/governance, services, and talent development. Korea's data policy has evolved in line with the development of ICT and national strategies over time. Currently, its focus is to strengthen future competitiveness centered on the data industry, which is a core resource in the era of the Fourth Industrial Revolution. Since mid-2020, there has been a marked shift toward AI-based services and industrial strategies grounded in data.

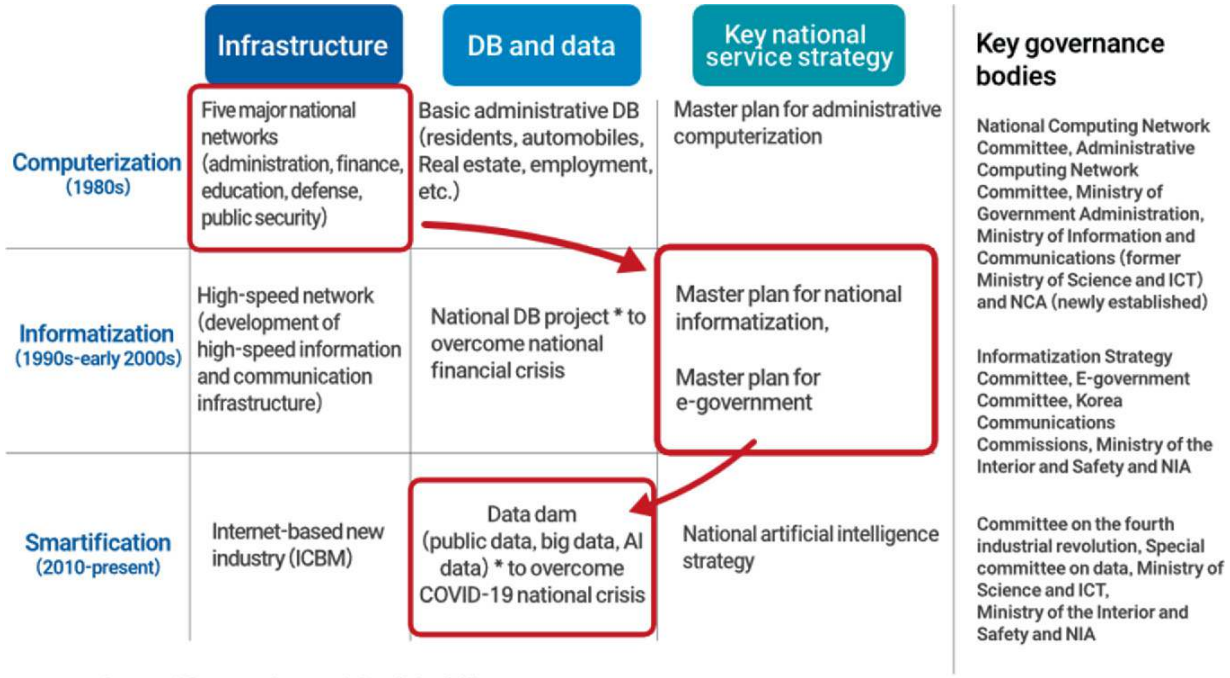


Fig. 1.1 Changes in the Korean Government's ICT policy

Korea's ICT governance has also undergone significant changes alongside its ICT policy. In the early stage of informatization, during the computerization phase, the Presidential Commission on National Backbone Computer Networks was established as a presidential advisory body, and the Ministry of Information and Communication (now the Ministry of Science and ICT) was created as a dedicated agency. To provide technical expertise, the government also established the National Computerization Agency (now the National Information Society Agency), thereby securing the professionalism and momentum needed to advance national informatization.

With informatization becoming a key national priority, the government initiated overarching ICT governance bodies such as the Informatization Strategy Committee and E-Government Committee. In particular, since the 2010s, with the full-scale promotion of AI and the data industry, the government has further refined its ICT governance structure by establishing entities such as the Presidential Committee on the Fourth Industrial Revolution and Special Committee on Data.

1.1.1 National Informatization and Construction of the National Backbone Computer Network

Korea's data policy began with the national informatization initiative. The government did not stop at simply building networks but strategically developed informatization by identifying and promoting core services deliverable through those networks as part of the national policy. In the process, the government secured the databases (DBs) and content required to develop a variety of services and created an environment conducive to their effective distribution. In other words, by developing key services and systems based on network infrastructure, the government naturally amassed vast amounts of data and databases.

From 1987 to 1996, the Korean government implemented the National Backbone Computer Network Project in two phases. This large-scale project was designed to improve the efficiency of national operations and accelerate socioeconomic development through the establishment of computing networks in five key areas: administration, finance, education/research, defense, and public security. Among these, the administrative computing network connected all central government ministries and local governments into a single network while also building core systems for national operations such as resident registration, real estate, and vehicle management. The administrative DBs required for these systems were also developed. Atotalofapproximately$5billionUSD,—an exceptionally large investment at the time—were allocated to the informatization project (Park & Park, 2003).

This bold investment by the government became a major driving force behind the growth of the ICT market. Companies such as KT and Dacom emerged to operate and maintain the national backbone networks. Large conglomerates such as Samsung, LG, and Daewoo actively entered the system integration (SI) market to win government contracts for data and service development. In addition, numerous SI startups were founded, allowing for the accumulation of broad

experience and expertise ranging from infrastructure construction to service development within the ICT sector.

Building on this foundation, the government launched the High-Speed National Information and Communications Network Project as a national initiative from 1995 to 2003, investing a total of 15.77 billion USD. By the time the project was completed in 2003, Korea had established the fastest ICT infrastructure in the world. During this period, the telecommunications equipment and service markets, as well as the high-speed network market, expanded rapidly. The total economic effect—including value-added generation, job creation, and production inducement—was approximately 7.3 times the original investment. Through this project, Korean network operators achieved substantial advancements in technology, infrastructure, and experience, reaching a level capable of competing with global companies.

Moreover, the government's proactive ICT policy contributed to segmenting the market into diverse sectors, thereby enhancing national ICT competitiveness. This included network operators such as KT, LGU+, and SKT; SI firms such as Samsung SDS, LG CNS, and SK C&C; and DB solution providers such as Altibase, Cubrid, and Tmax.

1.1.2 Transition to E-Government and the Data Economy

Entering the 2010s, the Korean government began shifting its focus from infrastructure- and network-centered informatization to software-centered development, placing greater emphasis on the importance of data. Building on the achievements of earlier informatization efforts, the government increasingly recognized data as a key strategic resource and began pursuing nationwide initiatives to actively promote its utilization. These efforts aimed to maintain Korea's status as an ICT powerhouse and strengthen its global leadership.

Recognizing the significance of data as a core national policy, the government released the "Big Data Master Plan for Realizing a Smart Nation" in 2012 (see Table 1.1). This plan outlines a strategy for managing the rapidly growing data at the national level while supporting the creative use of data in the private sector to enhance

future competitiveness. Under the National Informatization Strategy Committee, a sub-committee was established to coordinate and support core national policy agendas and identify forward-looking informatization themes. The government also announced the "Big Data Service Activation Strategy" to support the development of personalized and other data-based services, using massive datasets held by private sectors such as telecommunications, finance, and the Internet. These efforts were accompanied by initiatives focused on developing data platforms, fostering professional talent, and improving related systems to boost the data technology and industry ecosystem.

Table 1.1 Overview of the big data master plan and big data service activation strategy

Beginning in 2013, the Ministry of Science and ICT (formerly the Ministry of Science, ICT, and Future Planning) took the lead in establishing a data-driven industrial strategy. At the time, data, IoT, and cloud computing were identified as three core technologies for ICT-driven economic growth; however, Korea lacked global competitiveness in these areas. To address this, the government introduced the "Internet New Industry Promotion Plan" aimed at enhancing technological competitiveness in a short period and driving innovation across the ICT industry (Ministry of Science, 2013; see Table 1.2).

Table 1.2 Overview of the Internet New Industry Promotion Plan

In 2016, the era of AI was symbolized by the historic Go match between AlphaGo and professional player Lee Sedol, prompting the government to establish the Presidential Committee on the Fourth Industrial Revolution. With data, 5G networks, and AI designated as national priorities, systematic policy efforts began in earnest. This direction was further clarified in 2018 with the announcement of the "Data Industry Activation Strategy," which focused on building a comprehensive support system for the entire data lifecycle—from collection and construction to storage, distribution, analysis, and

utilization (Presidential Committee on the Fourth Industrial Revolution; See Table 1.3).

Table 1.3 Overview of the Data Industry Activation Strategy

Along with the efforts to activate the data industry and related services on a nationwide scale, the government also introduced a policy to fully open public data for broader use. The sustained investment under the 1990's slogan "Though we were late to industrialization, let's lead in informatization," has led Korea to consistently rank among the top countries since the mid-2000s in e-government evaluations conducted by international organizations such as the UN and OECD.

Based on significant investment and effort, the government has built a variety of e-government systems that have enabled the accumulation of vast amounts of administrative data. Moreover, the government has gone beyond merely accumulating administrative data, making this data accessible to the public to enhance administrative transparency and refining the data so that businesses can utilize it, thereby contributing to the activation of the national data economy.

1.2 Key Elements of Data Policy

The government's data policy can be broadly categorized into three key elements: (1) development of platforms and datasets, (2) strengthening of data policy governance and institutional foundations, and (3) cultivation of data professionals. These elements are organically linked to one another and together constitute the structure of the data policy.

1.2.1 Development of Data Platforms and Large-Scale Datasets

Korea's data policy aims to create an environment where public and private-sector data can be freely distributed, traded, and ultimately utilized. To achieve this, the government has actively intervened to

secure essential datasets early on, pursuing a strategy of building an ecosystem in which data can flow smoothly.

As illustrated in Fig. 1.2, the government has established the "Triangular Data Platform System," consisting of the Public Data Portal (open data platform), Big Data Platform (distribution platform), and AI Hub (AI platform). This system enables a virtuous cycle throughout the entire data lifecycle from collection and storage to distribution, exchange, and utilization.

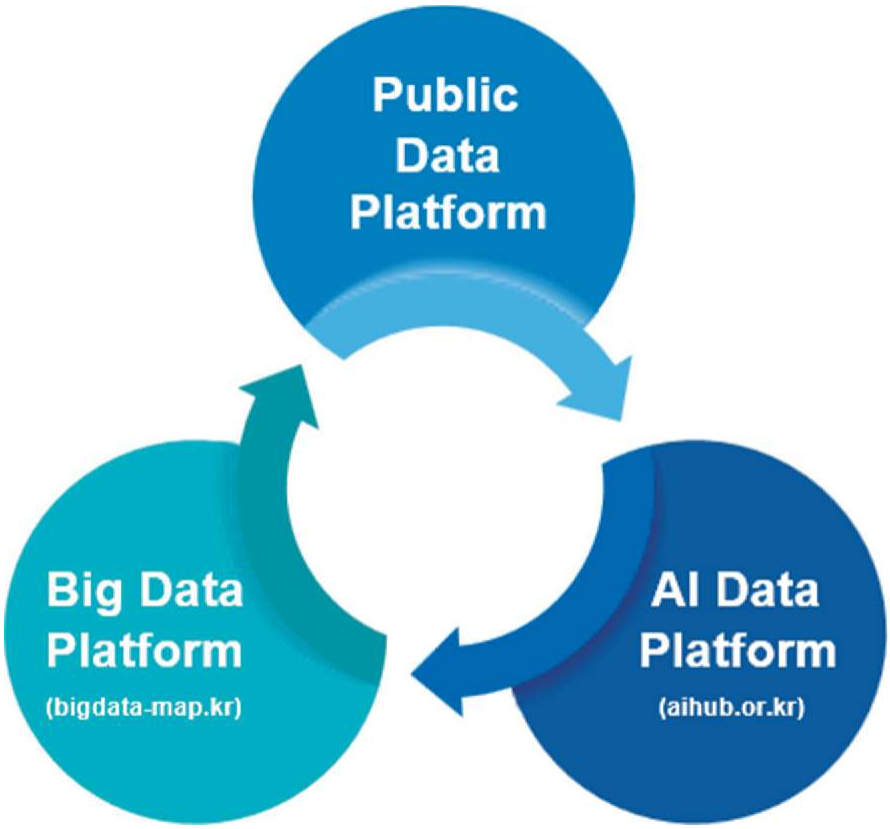


Fig. 1.2 Korea's data ecosystem. *Source* Compiled by the author

Through the development of this data ecosystem, the Korean government sought to expand existing data openness, secure essential data for the AI era, and establish platforms to enable the exchange and distribution of private-sector data, ultimately preparing for the era of data economy. In particular, the COVID-19 pandemic in 2020 served as a catalyst that accelerated the shift toward a data-driven economy. Rather than reducing investment in data-related initiatives during the

pandemic, the government decided to increase investment and accelerate its data economy strategy.

As part of the Digital New Deal, the government focused on building a "Data Dam" (see Fig. 1.3). The Data Dam is a large-scale data platform that collects, refines, and processes data generated from both public and private networks, making it available for businesses to utilize. The goal was to support companies in developing innovative services and in doing so, create numerous new jobs throughout the data generation and utilization process, thereby contributing to economic recovery amid the COVID-19 downturn. In this process, securing data-related professionals and creating digital jobs became key priorities.

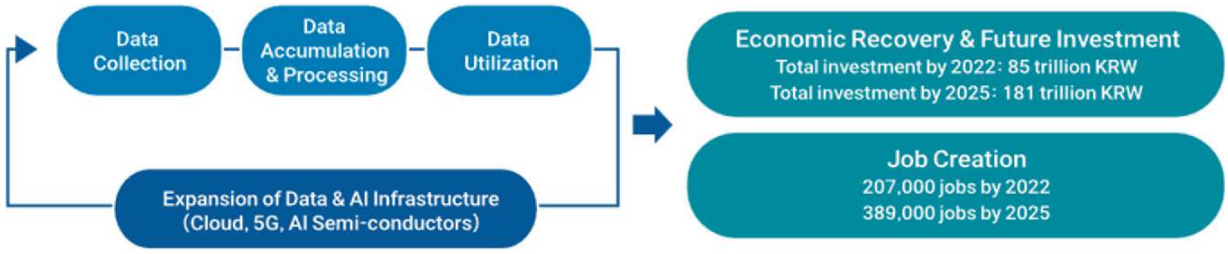


Fig. 1.3 Korea's Digital New Deal policy: The Data Dam. *Source*: Compiled by the author

1.2.2 Strengthening Data Policy Governance and Institutional Foundations

The governance structure for implementing data policy in Korea consists of a three-tiered organizational system. At the highest level is the Presidential Special Committee on Data, serving as the control tower overseeing all national data-related policies and initiatives. This committee determines the overall direction of National Data Policy and coordinates data-related projects and issues among government agencies.

The second tier consists of the Ministry of the Interior and Safety (MOIS) and the Ministry of Science and ICT (MSIT), which are responsible for formulating data-related policies and proposing decisions to the Special Committee, and initiating and managing new data projects. MOIS is in charge of all the policies and programs related to the public data produced and managed by central ministries, local

governments, and public institutions. MSIT is responsible for all the policies and initiatives concerning private-sector data generated by businesses.

The third tier includes the National Information Society Agency (NIA), which supports both ministries in the practical execution of their data policies and programs. By providing support to public and private-sector agencies, NIA plays a crucial role in bridging gaps and preventing conflicts between these policies.

To support the implementation of data policy, the government has also strengthened its institutional foundations. For example, in 2020, the Framework Act on National Informatization was comprehensively revised as the Framework Act on Intelligent Informatization, establishing a legal foundation for data-related policies. This revision aimed to build a virtuous-cycle ecosystem for the production, collection, distribution, and utilization of data in preparation for the Fourth Industrial Revolution. It also enabled the integrated management and support of public- and private-sector data and laid the legal groundwork for standardization to ensure interoperability between datasets.

In addition, to promote the public-interest use of private-sector data with public-good characteristics and foster the opening and distribution of such data, the government designated the NIA as the official support institution for efficient data management and collaboration (see Table 1.4).

Table 1.4 Key contents of the Data Industry Promotion Act

In 2021, based on previous experience with data policy, the government enacted the Framework Act on Data to advance the realization of data economy. This legislation, established to promote the activation and utilization of the data market, encompasses the entire data industry, from production, analysis, combination, and utilization of data to workforce development and international cooperation. It

explicitly recognizes the asset value of data and guarantees data-related rights.

The law also includes various mechanisms to encourage private-sector participation, such as the establishment of a framework for assessing data quality and value, training of professional data brokers, creation of data exchanges, and operation of a dispute mediation committee. These provisions are designed to address issues such as unauthorized acquisition, use, and disclosure of data, while also laying the foundation for a comprehensive national governance system (see Table 1.5).

Table 1.5 National Data Policy's implementation system

1.2.3 Data Talent Development Policy

Recognizing that data would become a core national asset in the era of the Fourth Industrial Revolution, the Korean government formulated and announced the "Data Industry Activation Strategy" in June 2018. This strategy reflects the growing importance of data as a new source of competitiveness that goes beyond traditional capital and labor, driving advancement and innovation across all sectors of industry.

At the time, South Korea faced several challenges in its data policy environment, including strict personal data protection regulations, a lack of high-quality data, a shortage of skilled professionals, and an underdeveloped foundation in data-related technologies and industries. In particular, demand for big data analytics continued to rise despite a significant shortage of experts capable of meeting this demand (see Fig. 1.4). According to the Software Policy & Research Institute (SPRi), between 2018 and 2022, the estimated demand for software professionals in the big data sector was 17,073, whereas the available supply was projected as 14,288, falling short by 2,785 individuals (Lee et al., 2018).

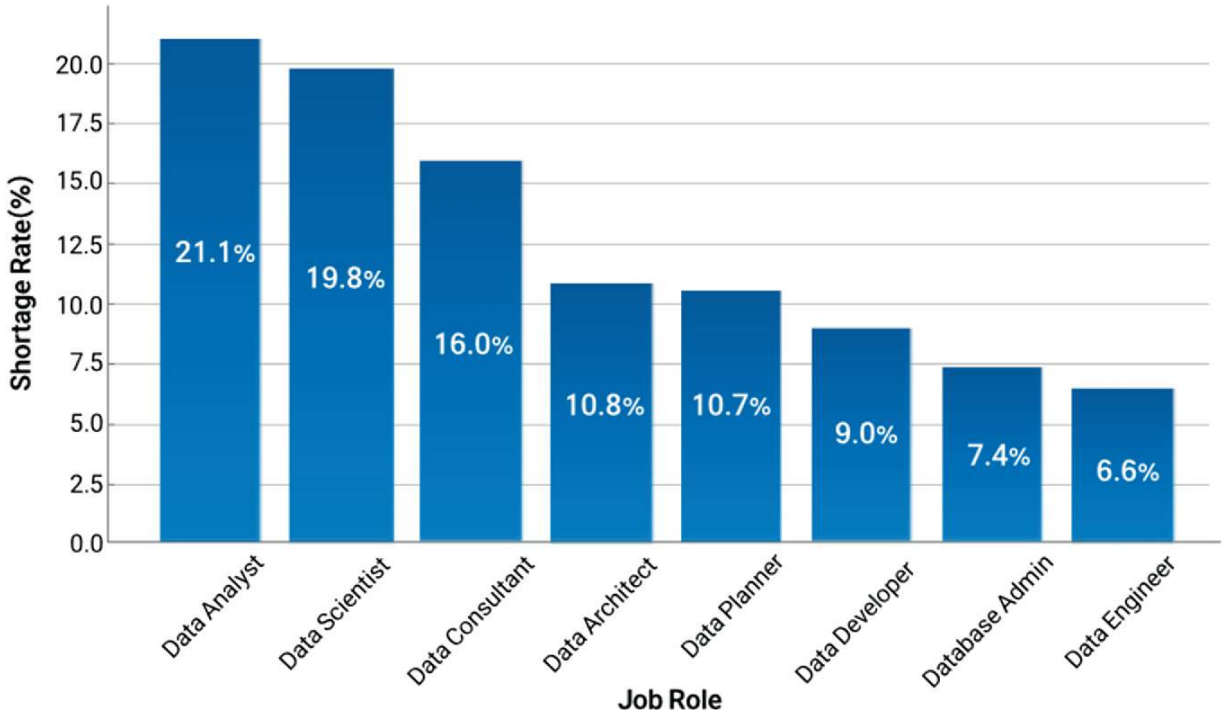


Fig. 1.4 Shortage rate of big data professionals by job category (2018–2023). *Source* Ministry of Science and ICT (2019)

In response to this gap, the Korean government launched the Data Industry Promotion Strategy, setting a national goal of training 50,000 young and skilled professionals in the data sector. As illustrated in Fig. 1.5, this initiative is structured around three key programs: first, fostering software talent through university-based education at SW-centered universities; second, intensively training data scientists through university-affiliated research centers known as Information Technology Research Centers (ITRCs); and third, strengthening practical training programs for employees already working in the industry.

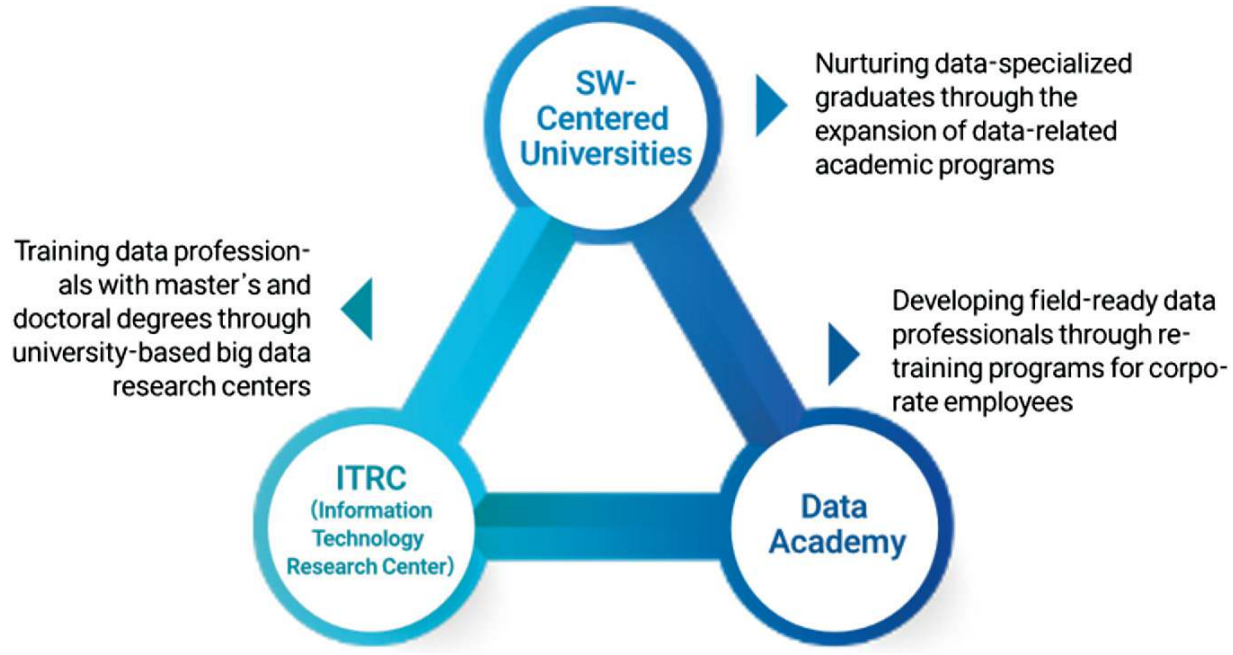


Fig. 1.5 Korea's data talent development system. Source: Presidential Committee on the Fourth Industrial Revolution (2018).

1.2.3.1 Nurturing Software Talent Through University-Centered SW Education

Since 2015, the government has designated and operated software-centered universities (SW Universities) to foster specialized software talent. These universities aim to cultivate industry-ready global professionals among software majors and convergence-oriented talent among non-majors through demand-driven education. As a result of this initiative, the number of software-related courses increased from 503 in 2015 to 4,053 in 2020, and the number of graduates specializing in software grew from 889 to 4,918. Building on this foundation, the government is now supporting the development of advanced professionals in data analytics.

1.2.3.2 Fostering Data Scientists Through the Information Technology Research Center (ITRC)

Based on the University ICT Research Center (ITRC) program, implemented by the government since 2000 to cultivate advanced ICT research personnel, the number of specialized big data research centers

expanded from one in 2017 to six by 2022. This program has contributed to the development of high-level ICT professionals by organizing outstanding research personnel within universities, supporting strategic technology R&D, and providing research funding, including stipends for master's and doctoral students.

1.2.3.3 Strengthening Practical Training for Industry Professionals

To meet industry demand, the Big Data Academy was established in 2013 to provide practical, job-oriented training for professionals with field experience, enabling them to grow into experts in the data field. The academy initially offered three programs—Big Data Planning Specialist, Big Data Technology Specialist, and Big Data Analysis Specialist—and more recently added Data Visualization Specialist, Data Brokerage Specialist, and Data Processing Specialist in response to emerging technologies and evolving corporate needs. By 2020, a total of 86 training sessions had been conducted, producing 2,073 data experts and supporting 398 pilot projects aimed at solving real-world business challenges (National Information Society Agency, n.d.).

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2026
S.-J. Eom et al., *Enabling Data-Driven Innovation and AI Government in Korea*
https://doi.org/10.1007/978-981-95-9175-6_2

2. The Three Major Data Platforms

Seok-Jin Eom[1] ✉, Yoon-seok Ko[2] ✉, Jeong Eun Park[3] ✉ and Keon Chul Park[4] ✉

(1) Graduate School of Public Administration, Seoul National University, Seoul, Korea (Republic of)
(2) National Information Society Agency, Daegu, Korea (Republic of)
(3) National Information Society Agency, Daegu, Korea (Republic of)
(4) Department of Business Administration, Tech University of Korea, Siheung-si, Korea (Republic of)

✉ **Seok-Jin Eom (Corresponding author)**
Email: sjum21@snu.ac.kr

✉ **Yoon-seok Ko**
Email: ysko@nia.or.kr

✉ **Jeong Eun Park**
Email: jpark0070@gmail.com

✉ **Keon Chul Park**
Email: parkkc07@tukorea.ac.kr

Keywords Open Government Data Portal – Big Data Platform – AI Hub – Public Data Act – AI training dataset – Data economy

2.1 Overview

As presented in Chap. 1, the Korean government has actively promoted policies for data openness and utilization to drive digital innovation and revitalize the data economy. Accordingly, the government has sought to strengthen transparency and efficiency in its transition toward a data-based society.

This chapter examines the roles, functions, and institutional foundations of the three major data platforms established as part of these efforts: (1) the Open Government Data Portal (data.go.kr, launched in 2013), (2) the Big Data Platform (launched in 2018), and (3) the AI Hub (launched in 2020).

These platforms perform distinct roles based on the types of data they provide and their intended uses, serving as core infrastructure for data policy. The Open Government Data Portal functions as an integrated access point that enables the public and businesses to easily obtain and use data from public institutions. The Big Data Platform revitalizes data transactions and supports private-sector innovation through sector-specific data distribution and integrated services. The AI Hub contributes to the development and expansion of the AI industrial ecosystem by building and providing AI training datasets under public leadership in response to the advancement and growing societal adoption of AI technology.

The Open Government Data Portal primarily offers data collected and managed across various public sectors, including administration, environment, and transportation. The Big Data Platform distributes data with high industrial value in each sector to promote private-sector utilization. The AI Hub provides advanced, processed data—such as vision, speech, and text data—suitable for AI training.

Advancements in the technological environment continue to enhance these platforms, offering more user-friendly experiences and improved technical support. Cloud computing, AI-based data analysis technologies, and the adoption of international standard metadata systems play key roles in improving data quality and usability. Additionally, user-centered interfaces and various support programs have further strengthened data accessibility and utilization.

2.2 Open Government Data Portal

2.2.1 Overview

The Open Government Data Portal is a national platform that provides integrated access to the open data held and managed by public institutions, in accordance with the Act on the Provision and Use of Public Data (Public Data Act). Through this portal, citizens can easily access and conveniently utilize the public data created or acquired by administrative and public institutions in various formats (file data, open APIs, visualizations, etc.) (see Fig. 2.1).



Fig. 2.1 Conceptual diagram of Open Government Data Portal.
Source National Information Society Agency (NIA) (2019a)

According to the Public Data Act, all public institutions, including central government agencies, local governments, and public organizations, are required to disclose the original data they possess. In principle, all data are subject to disclosure. However, exceptions may be made if disclosure can affect the safety of the public or the nation, with regard to personal information or matters of national security (see Table 2.1).

Table 2.1 Key provisions of the public data act

To prevent institutions from avoiding data disclosure, any request to exclude specific data from being made public must be submitted to the

Ministry of the Interior and Safety and reviewed and approved by the ministry's Open Data Strategy Committee. In addition, a formal procedure allows citizens to request the disclosure of non-public data, thereby establishing a strong legal obligation for public data disclosure.

According to the Public Data Act, the Ministry of the Interior and Safety must formulate a Basic Plan for the Promotion of Public Data Utilization every three years. Based on this plan, each institution must establish an annual implementation plan that specifies the content, method, and timing of data disclosure. The Ministry also evaluates the implementation of these plans as well as the overall status of public data openness and operation.

Public data policy is overseen by the Public Data Strategy Committee, and each public institution is required to appoint a Chief Officer for Public Data Provision to implement the policy. In addition, the Public Data Utilization Support Center (NIA) and the Dispute Mediation Committee play supporting roles in the implementation of public data policy (see Table 2.2).

Table 2.2 Key implementation framework of public data policy

2.2.2 Key Services Provided by the Public Data Portal

The Public Data Portal primarily provides three core functions: searching (data catalogs, national priority datasets, and the National Data Map), requesting (Data First Street, data provision requests), and utilizing (data visualization and analysis) public data (see Fig. 2.2).

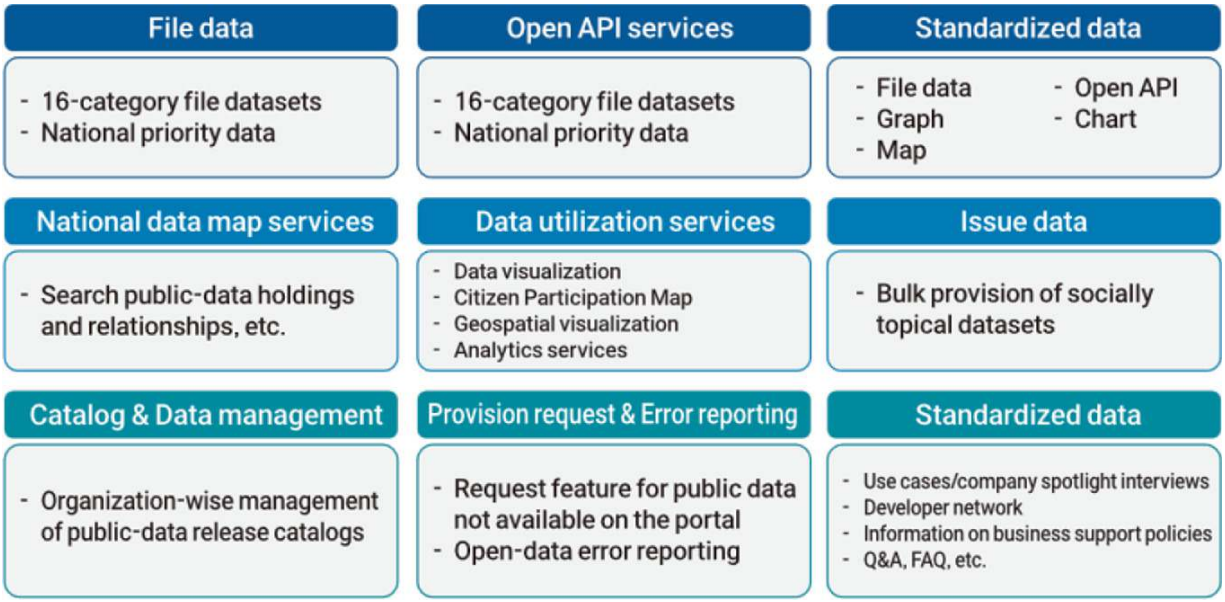


Fig. 2.2 Conceptual diagram of Open Government Data Portal.
Source National Information Society Agency (NIA), Republic of Korea (2019b)

Initially, the portal served as an integrated gateway for public data across government agencies, enabling unified access. However, with the advancement of the data economy and increasing public interest in data usage, the role of the Public Data Portal has evolved into a participatory platform that enhances the accessibility and usability of public data and supports more effective data openness.

To this end, the portal has strengthened its "user participation service" function by establishing a public data openness and utilization framework that enables users to create and share various outputs—such as products and services—using public data. In response to the global trend of digital transformation, the portal is also working to enhance the effectiveness of public data openness by establishing a demand-driven framework. This includes discovering and providing new datasets, improving the quality and sophistication of existing datasets, offering user-centered services, and enhancing user convenience features, thereby improving the overall usability and accessibility of public data.

In addition to opening and providing data, the Open Government Data Portal continues to strengthen its "user participation service"

function to promote active data utilization. This function supports users in creating and sharing products, services, and other outputs derived from public data. The portal also works continuously to improve data quality and discover new datasets. Users can directly download data or receive continuous updates through open application programming interfaces (APIs). Additionally, tools such as development toolkits and startup support programs are available to facilitate public data use. Visual resources such as the "National Data Map" shown in Fig. 2.3 further support data utilization by helping users easily locate the data they need.

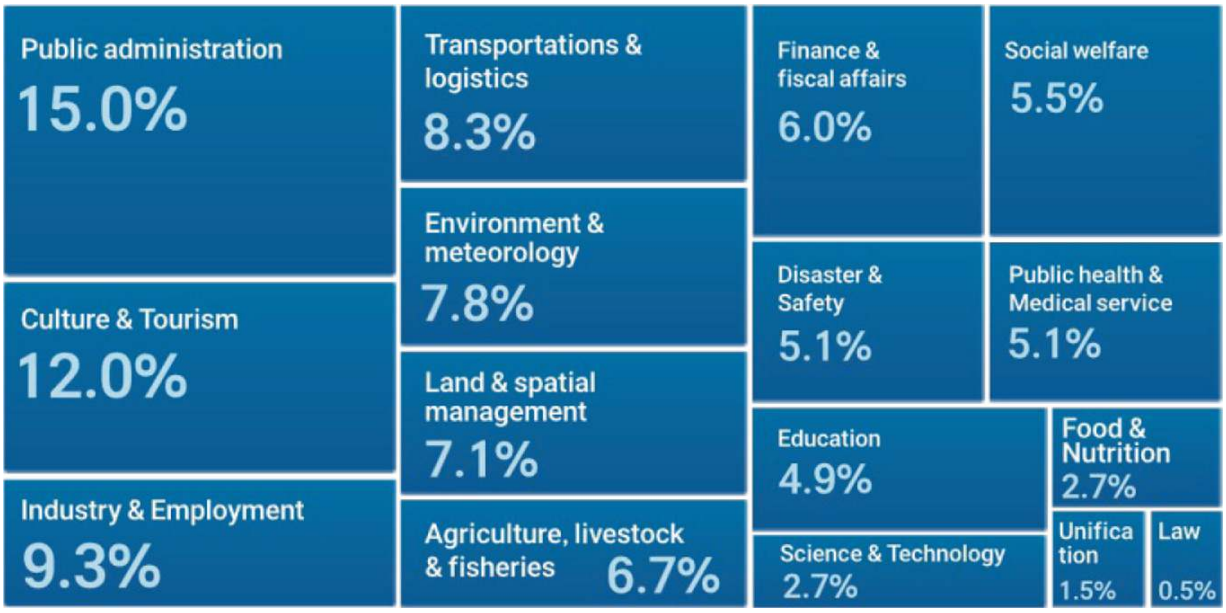


Fig. 2.3 National Data Map.
Source Open Government Data Portal website

2.2.3 Technological Foundation of the Open Government Data Portal

2.2.3.1 Data Production, Collection, and Processing

Open data are selected from datasets either produced or purchased by each public institution. Through services provided on the platform, users can preview the structure and partial contents of the data and perform basic processing. The data are made available in machine-readable open formats. An open format refers to a machine-readable

data format that can be processed (e.g., modified or edited) by at least one free, open-source software program without cost or usage restrictions.

The form of an open format may vary depending on the data type, such as images, text, tables, or geospatial data, and unstructured data must comply with international standards. Open formats are categorized into five levels, with representative formats at each level (see Fig. 2.4).

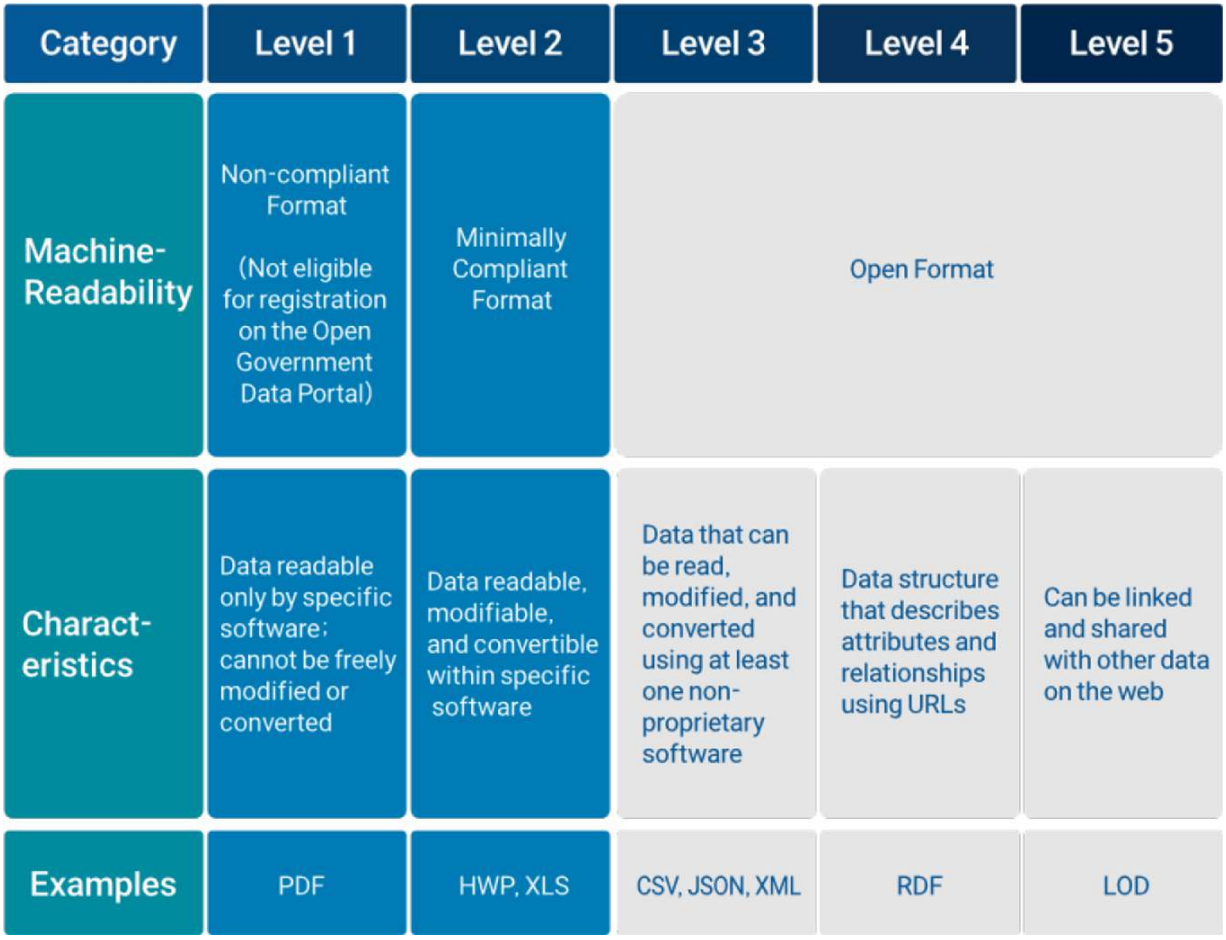


Fig. 2.4 Comparison of format levels for machine-readable data.
Source Ministry of the Interior and Safety (2021)

① Comma Separated Values (CSV): CSV is a format in which each item or piece of content is recorded and separated by a comma. CSV files are stored as text files and can be viewed or edited using word processors or text editors.

② JavaScript Object Notation (JSON): JSON is a widely used format for data exchange that follows the syntax of the JavaScript language. Since JSON is language-independent, it can be used across various programming languages such as C, C++, C#, Java, JavaScript, and Perl, allowing developers to work with it in the language of their choice.

③ eXtensible Markup Language (XML): XML is a markup language similar to Hypertext Markup Language (HTML). Designed to support the transmission of structured documents over the web, XML ensures the independence of each component within a document. This provides advantages such as compatibility, content independence, and ease of modifying individual elements. Unlike HTML documents, which are typically stored as a single file, XML documents can be structured with separate files for individual elements, allowing efficient storage, retrieval, and reuse. XML also facilitates document exchange across different systems, ensuring compatibility even when different applications are used.

④ Resource Description Framework (RDF): RDF is a general semantic web technology language used to describe metadata on the web, often using XML or similar formats. By defining the properties (attributes) of metadata, RDF enables efficient data exchange between applications and supports interoperability across systems.

⑤ Linked Open Data (LOD): LOD is a concrete implementation method of the “Data Web,” a concept promoted by the Semantic Web. It involves providing structured data in RDF format using RESTful protocols on websites. Linked Data enables diverse and related datasets to be interconnected across the web, allowing them to function collectively as a unified knowledge base.

⑥ Open Document Format (ODF): ODF is an office document file format based on an XML file structure. It is supported by a wide range of open-source and commercial office software, including

Hangul Word Processor (HWP), Microsoft Word (MS Word), and LibreOffice. The file extension for word processing documents is.odt, and most office software allows users to save files in this

format using the “Save As” function. Other common extensions include.ods for spreadsheets and.odp for presentations.

2.2.4 Data Registration and Provision Methods

Datasets designated for public release are registered on the Open Government Data Portal and provided through three main methods: downloads, open APIs, and LOD. Open APIs are suitable for large-scale data, whereas LOD is appropriate for static data with high interconnectivity.

As illustrated in Fig. 2.5, the processes of data collection, creation, processing, and operation occur within the internal systems of each institution. However, data registration management and post-release management—being user-facing functions—are handled through the institution’s external systems. Since public data are open and accessed by the private sector via the Open Government Data Portal, separate security compliance regulations are applied to internal and external systems.

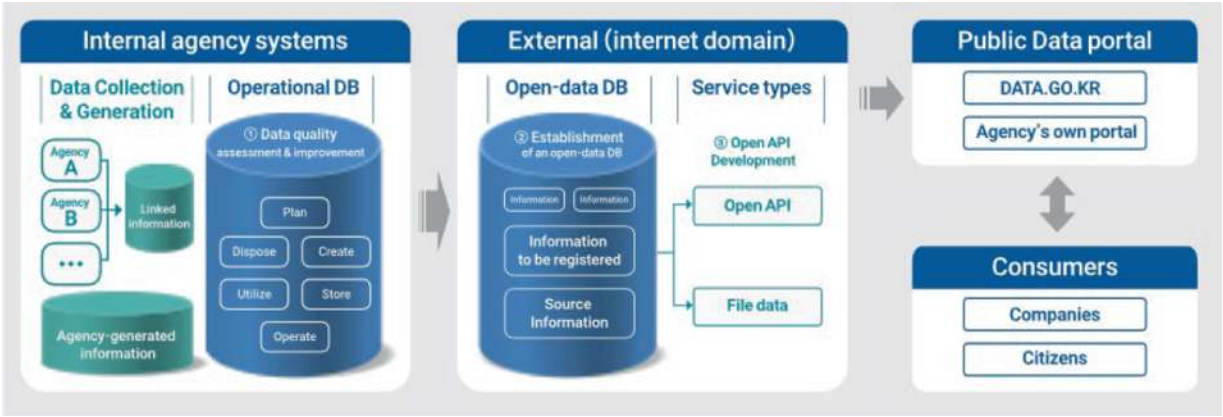


Fig. 2.5 Open Government Data Openness Framework.
Source Ministry of the Interior and Safety (2021)

The technological foundation of the Open Government Data Portal is built on an open standard structure that enables seamless

interconnection and efficient dataset management across institutions. This structure provides a common set of rules for dataset management and openness, ensuring interoperability and broad applicability.

Data are delivered through an API gateway, and a key-based authentication system is used to ensure security. The system also automates API request validation and performs regular health checks. When API errors occur, the system promptly notifies responsible personnel to enable an immediate response.

As of October 2024, the Open Government Data Portal operates on both government cloud (G-Cloud) and a private cloud infrastructure. However, the combined use of distributed and centralized open API integration models has resulted in inefficiencies in API management and failure response.

To resolve these issues, the portal is prioritizing the migration and redesign of key open APIs with high private-sector demand to a cloud-native environment. The objective is to establish a hybrid open API system that integrates G-Cloud and private cloud infrastructures using open-source technologies. This system will support both centralized and distributed API environments to ensure stable operation, efficient management, and reliable technical support.

Additionally, the portal plans to support the development, registration, and standardization of new open APIs, including those intended for use in national emergency or disaster situations (see Fig. 2.6).

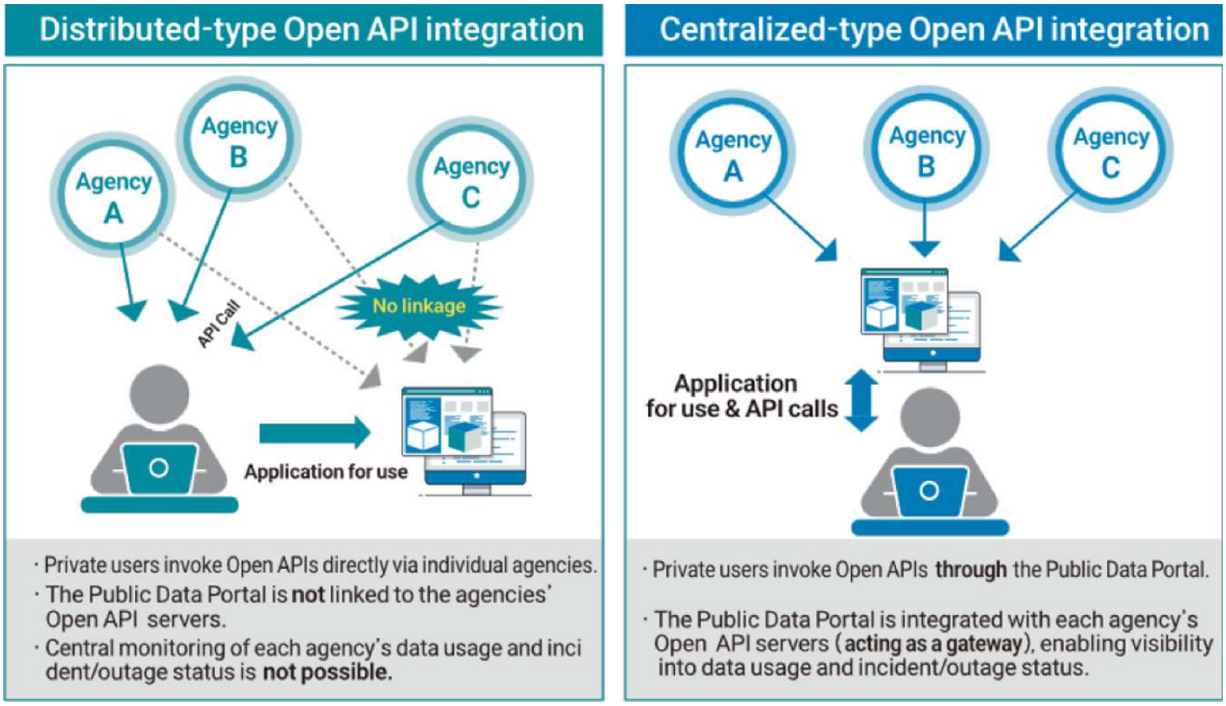


Fig. 2.6 Open API integration strategy based on multi-cloud (public and private) infrastructure. *Source* National Information Society Agency (NIA), Republic of Korea (2024b)

2.2.5 Metadata Management and Data Integration

Metadata management within the Open Government Data Portal serves as a core element in implementing an integrated management framework for Open Government Data. The government, by applying a metadata management system in central administrative agencies, local governments, and public institutions, established a structure for collecting and managing the metadata held and released by each organization.

Additionally, efforts have been made to standardize metadata and ensure data quality by managing a standardized (data) terminology dictionary. Building on this foundation, the government expanded the integrated data management platform, to develop a shared data registration and management system for the registration, storage, and interconnection of shared-use data across other data systems.

Based on these efforts, the government is promoting a government-wide common system that supports the provision, linkage, and joint utilization of data, along with data analytics, through integration of the

Open Government Data Portal, central metadata repository, and shared-use datasets.

2.2.6 Development of the Next-Generation Open Government Data Portal

The government is currently promoting the development of the next-generation Open Government Data Portal, to establish an efficient public data provision system and strengthen data integration and utilization. This initiative aims to resolve inefficiencies in the current data management and operation practices in building a government-wide integrated data provision system (One Portal).

The next-generation Open Government Data Portal development project is broadly divided into three components: establishing a Digital One Platform, service enhancement, and the design and construction of a multi-cloud-based infrastructure.

By establishing the Digital One Platform, the government plans to centralize the integrated provision and management system of government-wide data and open APIs. In service enhancement, the current keyword-based search service will be improved by adopting deep learning technologies, such as natural language-based semantic search, to allow users to explore the desired data more intuitively and accurately. Additionally, the recommendation service will be upgraded with new algorithms based on collaborative and content-based filtering methods. A National Data Map service will be designed to offer diverse large-scale datasets with associated visualizations through data integration for user convenience.

Finally, by designing and building a multi-cloud-based infrastructure, the current services operated under a unified database will be modularized and transitioned to individual database structures.

2.3 Big Data Platform

2.3.1 Overview

The Big Data Platform is a data distribution and utilization system established to enable the collection, storage, analysis, and exchange of data from public and private sources. To facilitate smooth data distribution, the government supports both data providers (companies and institutions) that supply the data to the platform and data centers that operate and distribute the data, thereby assisting data users (such as businesses and researchers) in utilizing the data effectively.

Initially, the Big Data Platform focused primarily on functions related to data trading and distribution. However, its functionality has since expanded to support direct data analysis and utilization. Consequently, the platform has evolved beyond a mere data marketplace into a data ecosystem involving data providers, users, and intermediaries. It now serves as a venue for generating new insights and business value (see Fig. 2.7).

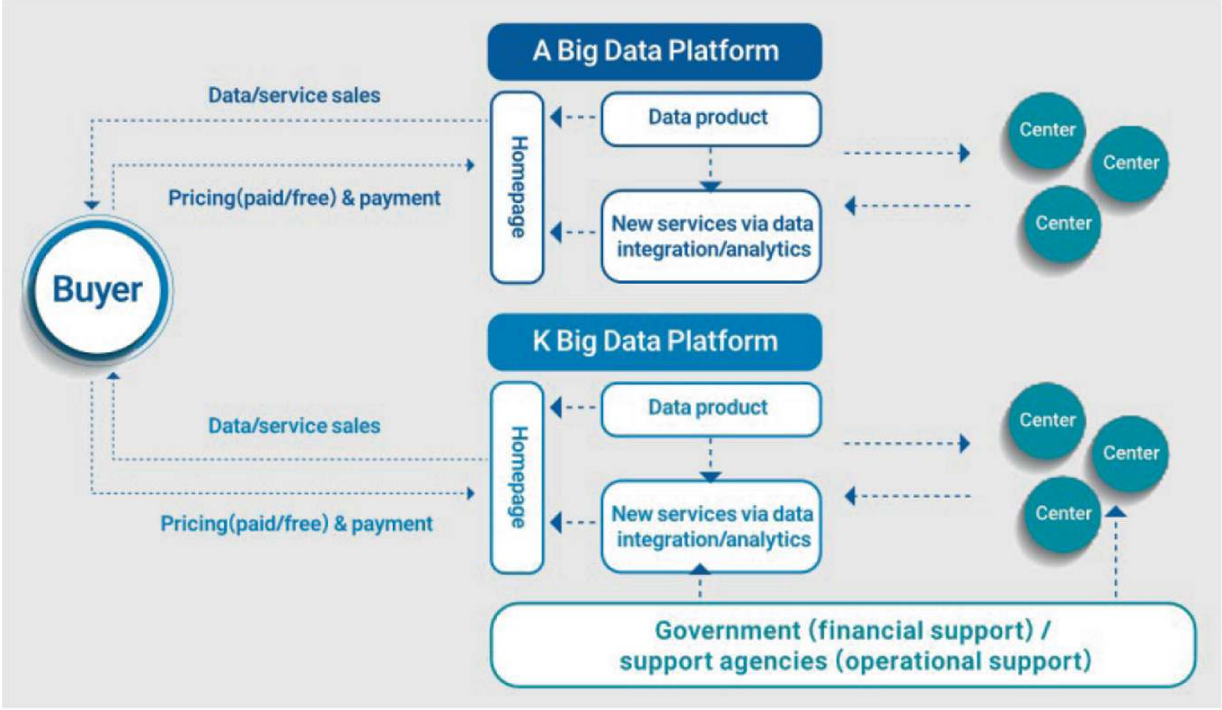


Fig. 2.7 Structure and operational framework of the Big Data Platform. *Source* Jeong, Junhwa (2020)

The Big Data Platform offers an open distribution and transaction environment where demand-side users can purchase data products

that have been collected, processed, and analyzed according to industrial demand. As shown in Table 2.3, as of March 2024, a total of 21 domain-specific platforms in telecommunications, finance, environment, and welfare, encompassing approximately 230 centers, have been established.

Table 2.3 Structure and data products of the 21 sectoral Big Data Platforms

Users can access each sector's platform via an Integrated Data Map, which allows easy search and utilization of big data across platforms from a single access point. Currently, the platform is also promoting services that support data-driven competitiveness among businesses, to bridge the data utilization gap between large and small enterprises. As of December 2024, the platform has distributed 74,436 file datasets, 11,813 open APIs, and 11,010 standardized datasets.

2.3.2 Key Services Provided by the Big Data Platform

The core function of the Big Data Platform is to provide an open transaction environment where data consumers can search for and purchase data products. After completing a simple registration process, which is open to everyone, users can purchase the desired data products on the platform, either free of charge or for a fee.

The platform offers keyword-based search functionality that allows users to search for data products by category and keyword and view detailed information about each product. These data products are supplied by individual data centers, and purchasers can either download the data directly or access it via an API through the platform.

However, depending on the nature of the data, some platforms impose restrictions on membership and data usage conditions (see Fig. 2.8). These restrictions are necessary to ensure data security and ethical use. For example, the small and medium enterprise (SME) platform limits membership to corporate users, focusing on providing customized data services for SMEs. This approach is intended to

strengthen SMEs' data utilization capabilities and prevent misuse by individuals or large corporations.

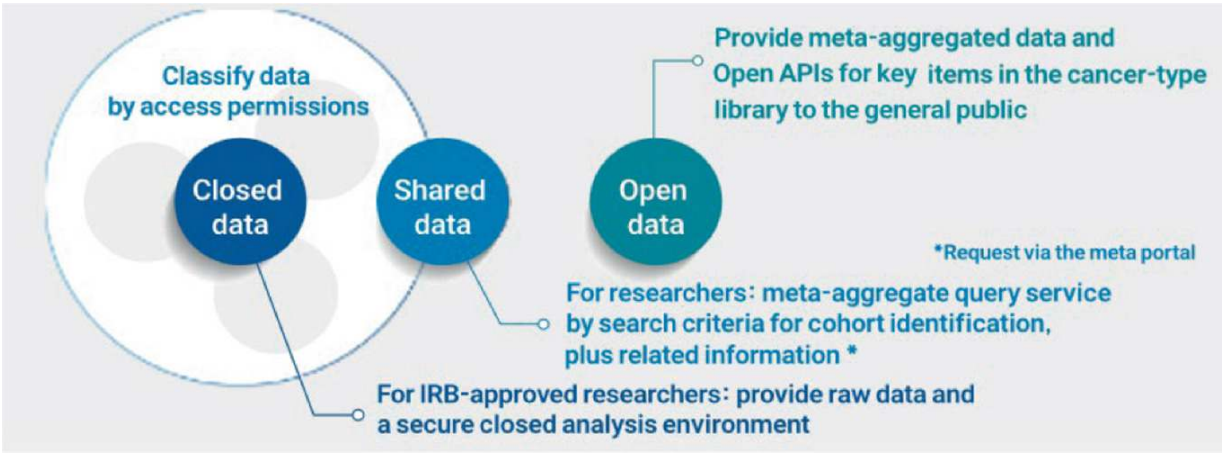


Fig. 2.8 Usage permissions by data type in the Big Data Platform.
Source Big Data Platform Website (www.bigdata-cancer.kr)

In addition, the Healthcare Platform includes sensitive patient information, and must therefore comply with the Personal Information Protection Act and bioethics-related regulations. As a result, certain raw datasets are only available to researchers or institutions that have received approval from an Institutional Review Board (IRB).

The Big Data Platform operates the "Data 114" service, which connects users with data providers, serving as a mediator for companies that can process and deliver the required datasets. Additionally, it runs a "Transaction Support Center" to support the policies related to data trading and applies a standardized data pricing model to manage transactions systematically.

One of the platform's distinctive services is its data integration (or data fusion) function. This enables the creation of new data products by combining heterogeneous datasets. For example, by integrating weather and credit card transaction data, the platform can provide insights into weather-based shopping trends. Similarly, on the Coastal Big Data Platform, data on coastal weather, climate, and topography are combined to predict coastal disasters and analyze potential damage scales (see Fig. 2.9).



Fig. 2.9 "City in Crisis" screen of the Coastal Big Data Platform's Safety Service.
Source Coastal Big Data Platform Website (www.bigdata-coast.kr)

The Big Data Platform supports an Integrated Data Map that enables users to search for and utilize data from both the public and private sectors in one place. As of April 2025, the Integrated Data Map provides a comprehensive view of data produced by 21 sectoral platforms, 230 data centers, and services such as AI Hub (see Fig. 2.10).



Fig. 2.10 Integrated Data Map homepage.
Source Integrated Data Map homepage (www.bigdata-map.kr)

Through this map, users can conveniently search for desired datasets using keywords and advanced search features and visually grasp the overall status of available data. The platform also offers data profiling and data curation services to help users verify data accuracy and identify potential issues.

The Big Data Platform also operates a Big Data Platform-Based Analytics Support Program (see Fig. 2.11). In 2022, 20 demand-side companies and six analytics firms participated, and in 2023, participation increased to 49 demand-side companies and 19 analytics firms.

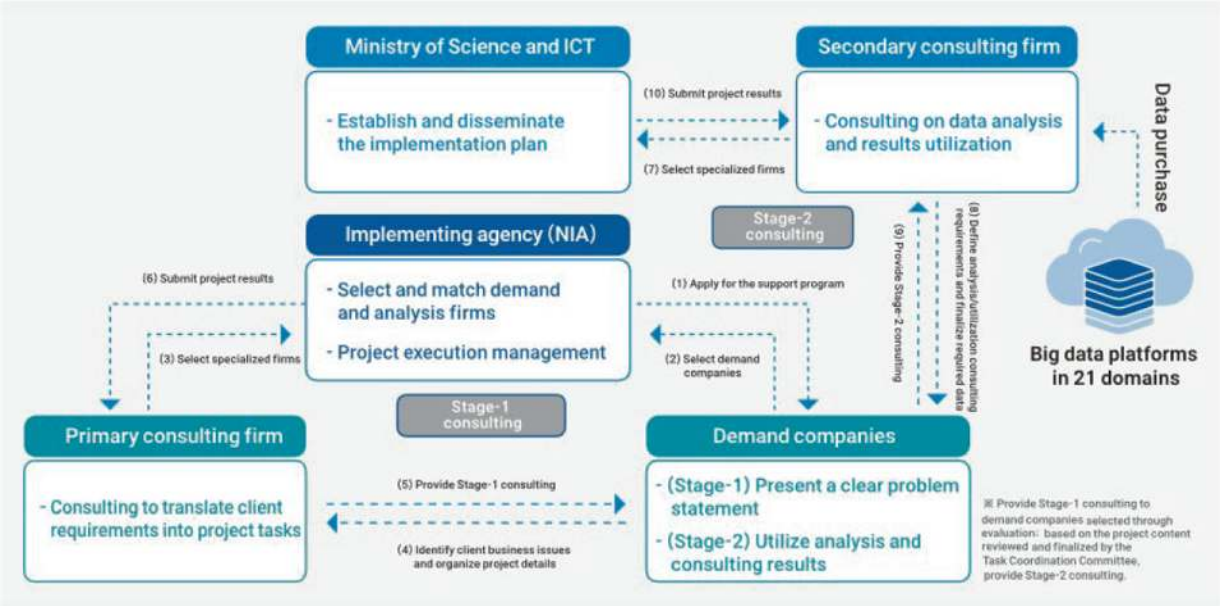


Fig. 2.11 Overview of Big Data Platform-based analytics support program.
Source National Information Society Agency (NIA), Republic of Korea (2024d)

The Big Data Platform plans to expand the scale of this support program and establish a data analysis and consulting support system for client companies and domain-specific analytics firms, to collaborate and drive business innovation for small and venture enterprises.

The project consists of two phases of consulting. Phase 1 involves consulting for planning analytical tasks and establishing a support

strategy. Through in-depth interviews conducted by professional consulting firms, this phase identifies the business issues of the client company, uses data to clarify the problems to be addressed, and defines specific requirements to derive analytical tasks.

Phase 2 identifies both the data the client company already possesses and any additional data required based on the results of Phase 1. Through data analysis and utilization consulting, it provides solutions to business challenges and presents mid- to long-term data utilization strategies.

This project supports the development of strategies to enhance corporate competitiveness through data by matching a wide range of client companies with specialized analytics firms, aiming to resolve the data scarcity issues faced by SMEs and promote business innovation through data use.

The project provides customized analysis and utilization consulting via domain-specific analytics firms to help SMEs address various real-world challenges and generate tangible outcomes. Furthermore, it contributes to establishing a sustainable data market cycle of supply and demand, helps companies discover new data-driven opportunities, and supports their growth.

2.3.3 Technological Foundation of the Big Data Platform

The infrastructure of the Big Data Platform is designed according to the characteristics of the data held by each individual platform and center, with standard guidelines provided to ensure interoperability with other platforms. A common infrastructure based on private cloud services is utilized to handle large-scale traffic and ensure security. Technical foundations such as the C*omprehensive Knowledge Archive Network (*CKAN) (Harvest), NiFi, and open APIs are applied to support data management, search, access, and distribution (see Fig. 2.12).

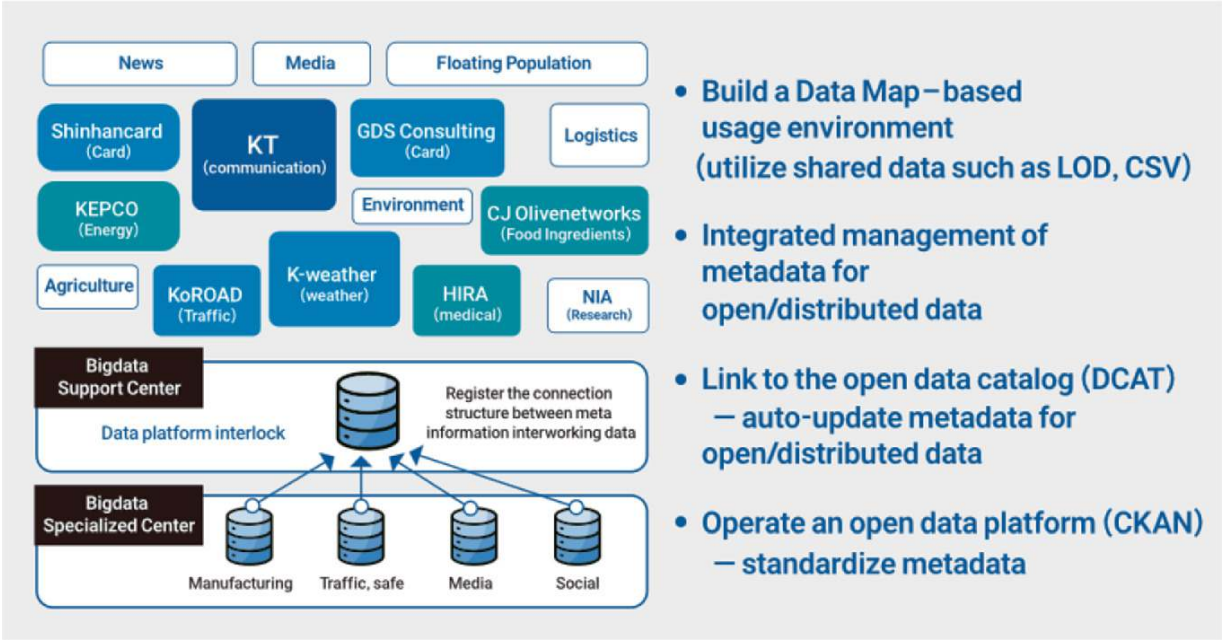


Fig. 2.12 Big data sharing and integration framework.
Source Statistics Korea, Statistical Education Center (2019)

Platforms built on a data hub model utilize CKAN and the Data Catalog Vocabulary (DCAT) standard to facilitate the opening and integration of public and private-sector data, and the construction of data maps helps users easily explore available datasets. Furthermore, by employing RDF and LOD, data interoperability is enhanced. To ensure data quality and reliability, object identifiers (OIDs) are applied, establishing a solid foundation for data distribution.

Platforms built on a data hub model utilize CKAN and the DCAT standard to support the opening and integration of public and private-sector data. These platforms also enable users to easily explore datasets through the construction of a data map.

In addition, data interoperability is enhanced by employing RDF and LOD. To ensure data quality and trustworthiness, the platforms implemented OID, to establish a robust foundation for data distribution and management.

2.4 AI Hub

2.4.1 Overview

As an integrated platform, AI Hub provides AI datasets, AI software APIs, and computing resources to support the development of AI technologies, products, and services. As the performance of artificial intelligence models is highly dependent on the quantity and quality of training data, securing large-scale, high-quality AI training datasets is essential. However, collecting and processing such data incur significant time and cost, creating a high barrier for startups, SMEs, and universities conducting AI research and development. Therefore, demand for AI training dataset expansion tailored to domestic context has been growing.

AI Hub publicly provides AI training datasets in 14 domains—Korean language, image and video, health care, transportation and logistics, disaster and safety, agriculture/livestock/fisheries, culture and tourism, sports, education, robotics, manufacturing, intellectual property, legal, and finance—developed through the "AI Training Dataset Project." Most of these datasets were constructed under government leadership. Moreover, high-value datasets accumulated at the national level are made available.

The AI Hub allows not only non-commercial use for research purposes but also commercial use for the development of products and services by private companies. To support this, AI Hub focuses on building and releasing AI training datasets that are technically and industrially promising and broadly applicable. The NIA, which oversees the operation of AI Hub, provides standard guidelines for each domain and implements a multi-stage quality verification process to ensure the reliability of the datasets (see Fig. 2.13).

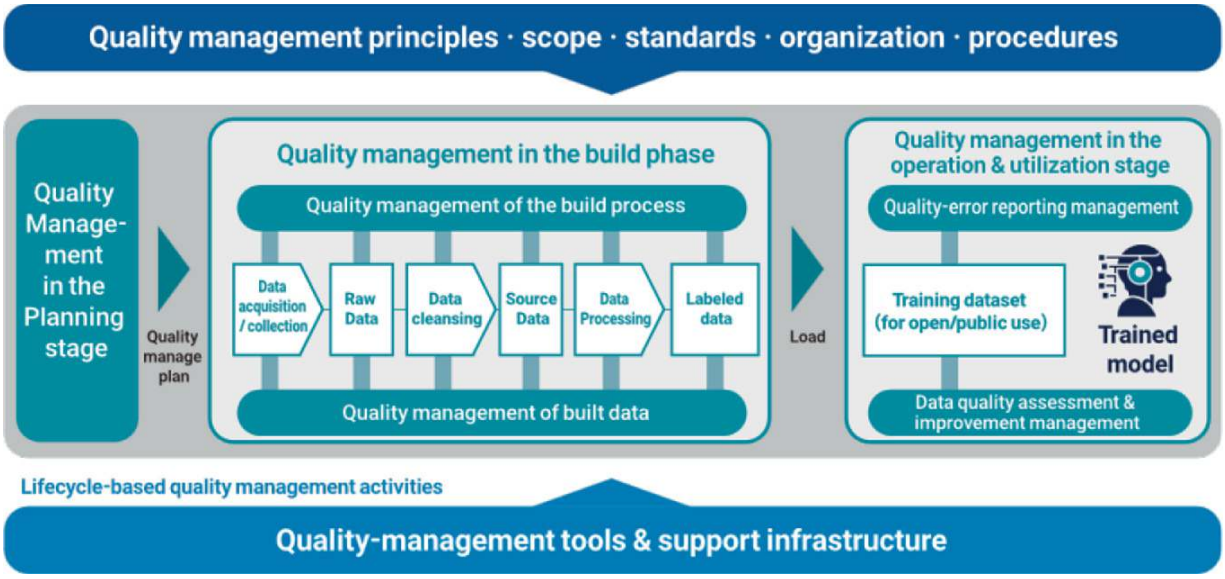


Fig. 2.13 Lifecycle-based quality management framework for AI training dataset project. *Source* National Information Society Agency (NIA), Republic of Korea (2023b)

The goal of AI Hub is to create a virtuous lifecycle ecosystem in which the public sector provides foundational data for AI technology development, the private sector uses these data to advance technologies, and the expanded datasets are then rereleased for further use.

In addition to datasets, AI Hub also provides data-labeling tools and AI service models, thereby supporting both the development of AI technologies and the dissemination of AI services. The government has allocated significant funding to the AI training data construction support program to ensure the availability of high-quality datasets. This program is also expected to generate employment opportunities by engaging labelers (crowd workers) in the data construction process.

2.4.2 Core Services Provided by AI Hub

The AI Hub provides a wide range of services to promote AI research and industry development, including data provision, computing support, and AI solution adoption assistance.

As of August 2025, AI Hub offers 914 types of AI training datasets across 14 domains. These datasets, labeled and constructed through the AI Training Dataset Project, are publicly available. AI Hub also

provides training datasets sourced from domestic and international organizations and companies. Each dataset includes raw data, labeled data, and metadata, along with training videos, example AI models, and source code to support effective use.

Beyond datasets, AI Hub offers services to support broader AI adoption. These include open APIs, AI educational content categorized by level, type, and year, and a Korean LLM leaderboard. Through these services, AI Hub functions as a comprehensive AI infrastructure platform that supports not only datasets but also open APIs, software technologies, and learning resources (see Fig. 2.14).

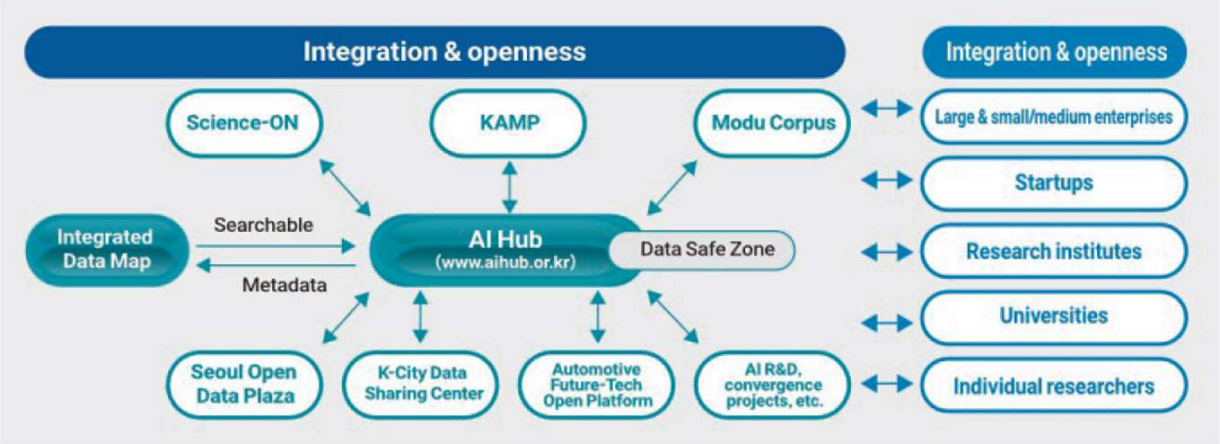


Fig. 2.14 AI Hub homepage interface.
Source AI Hub Website (www.aihub.or.kr)

Another key offering is the AI computing service, which provides high-performance computing (HPC) resources needed for AI R&D. These resources are leased from private cloud providers and made available to SMEs, startups, and universities, primarily in the form of graphics processing unit (GPU) servers for training AI models (see Fig. 2.15).

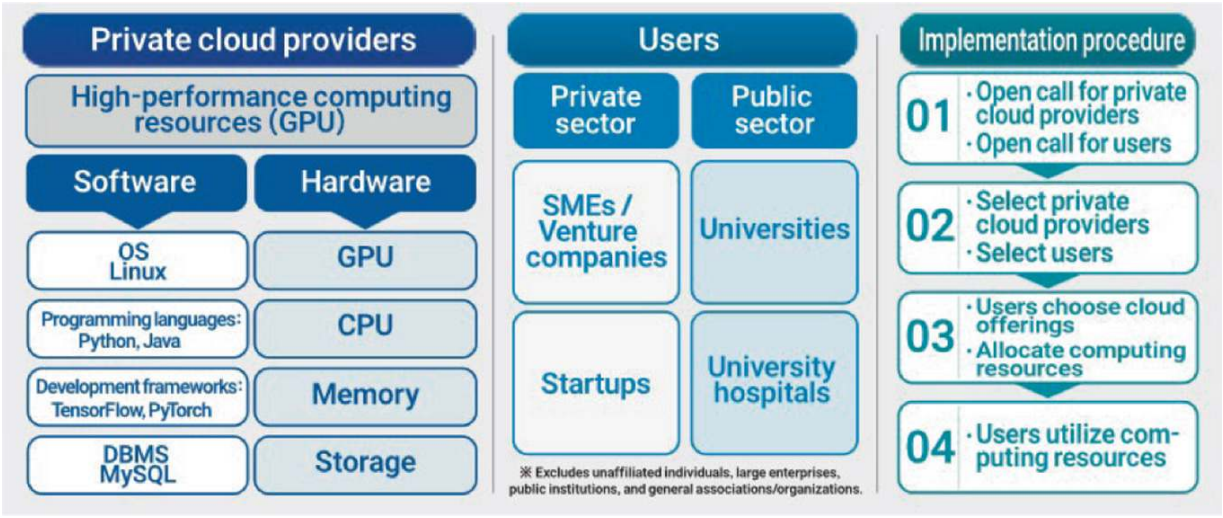


Fig. 2.15 AI computing service management framework of AI Hub.
Source AI Hub Website (www.aihub.or.kr)

Finally, the AI voucher program enables AI solution providers (SMEs and startups) to offer their AI solutions to demand-side companies—other SMEs or mid-sized firms—through a government-issued voucher. This program helps create new market opportunities and fosters the expansion of the AI industry ecosystem. As of January 2025, a total of 1,390 AI solutions were registered, spanning areas of intelligence such as analytics, visual, language and speech, and behavioral intelligence. These include a wide range of capabilities, from machine learning operations (MLOps) platforms and deep learning-based solutions to edge device-based embedded systems.

2.4.3 Technological Foundation of AI Hub

The AI Hub consists of AI training data construction, data integration and management systems, the portal system, and utilization services. It has also established a data governance and performance management system to provide high-quality AI data. Through open APIs, it connects with public and private institutions and centrally manages metadata.

The cloud infrastructure of AI Hub comprises the AI Hub Portal and AI Training Dataset Service. First, the AI Hub Portal focuses on core service functions such as AI data search, utilization, statistics, and error reporting. Additionally, to ensure the safety of data containing sensitive

information, it has an established “Sensitive Information Handling Policy” to provide a foundation for secure data use without concerns over data leakage or rights violations.

The portal includes subcomponents such as metadata, data quality, and storage management systems, sensitive information analysis, along with data linkage and governance functions. Furthermore, within the portal, role-based access control is implemented for users to access sensitive information only within permitted scopes. A “Safe Zone” service has been built to ensure this, allowing secure data usage during external collaboration by managing access logs and using encrypted transmission methods.

The AI Training Dataset Service applies object storage technology to support large-scale processing, convenient data search, and file access. Through this technology, it assigns identifiers to metadata for duplicate data and content lifecycle management, thereby facilitating data access and ensuring efficiency in unstructured data management. It also enables integration and linkage with other services provided within the cloud.

Additionally, through a user-friendly interface, users can monitor and control the upload and download of large datasets. The dataset version control feature supports dataset reuse during model training and allows restoration to specific points in time.

As shown in Fig. 2.16, AI Hub has introduced a “Safe Zone” service to enable the secure use of datasets containing sensitive information, such as those in the healthcare domain. The Safe Zone provides an environment where users can develop AI models without exporting data, in both online and offline settings.

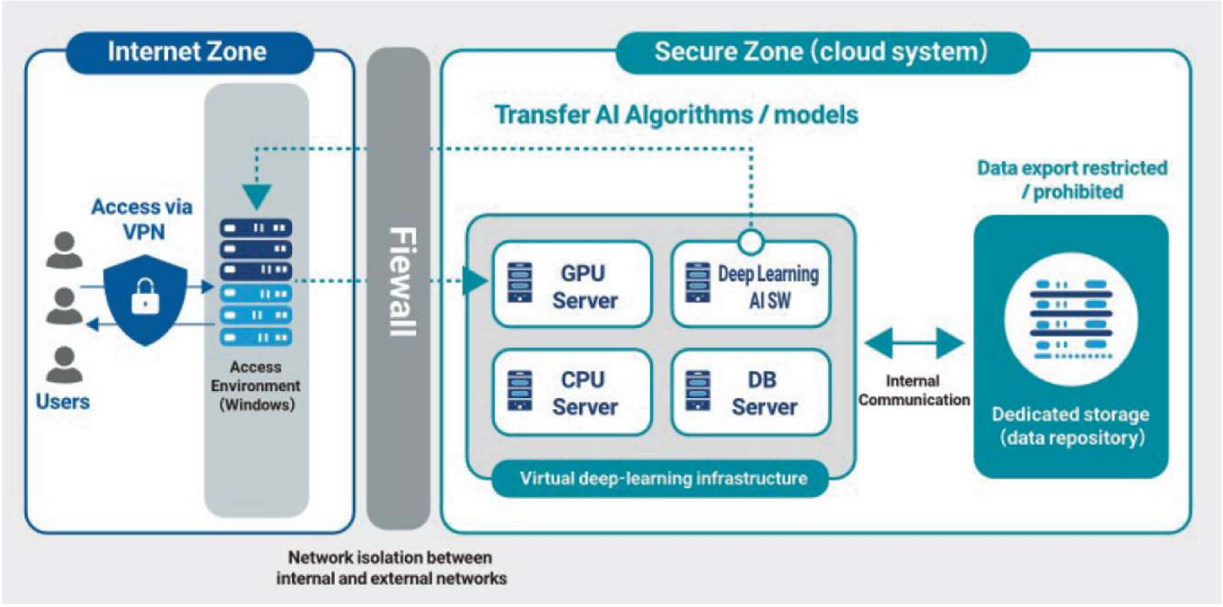


Fig. 2.16 Overview of AI Hub's Safe Zone.
Source AI Hub (n.d.)

The online Safe Zone allows users to access and analyze data from anywhere—such as their home, laboratory, or office—via a secure online network. In contrast, the offline Safe Zone is designed for highly sensitive data requiring additional security measures, enabling data access and analysis only within a designated physical space. This applies to healthcare data that includes facial features, skin, hair, and other high-sensitivity elements.

The Safe Zone separates external networks from internal infrastructure through virtual private network (VPN)-based access. It strictly controls data export and provides a virtual deep learning infrastructure for AI model development. This system allows users to train AI models in a secure environment, exporting the trained models and source code only when necessary, thereby ensuring that the sensitive data remain protected at all times.

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2026
S.-J. Eom et al., *Enabling Data-Driven Innovation and AI Government in Korea*
https://doi.org/10.1007/978-981-95-9175-6_3

3. Achievements and Challenges of Korea's Data Policy

Seok-Jin Eom[1] ✉, Yoon-seok Ko[2] ✉, Jeong Eun Park[3] ✉ and Keon Chul Park[4] ✉

(1) Graduate School of Public Administration, Seoul National University, Seoul, Korea (Republic of)
(2) National Information Society Agency, Daegu, Korea (Republic of)
(3) National Information Society Agency, Daegu, Korea (Republic of)
(4) Department of Business Administration, Tech University of Korea, Siheung-si, Korea (Republic of)

✉ **Seok-Jin Eom (Corresponding author)**
Email: sjum21@snu.ac.kr

✉ **Yoon-seok Ko**
Email: ysko@nia.or.kr

✉ **Jeong Eun Park**
Email: jpark0070@gmail.com

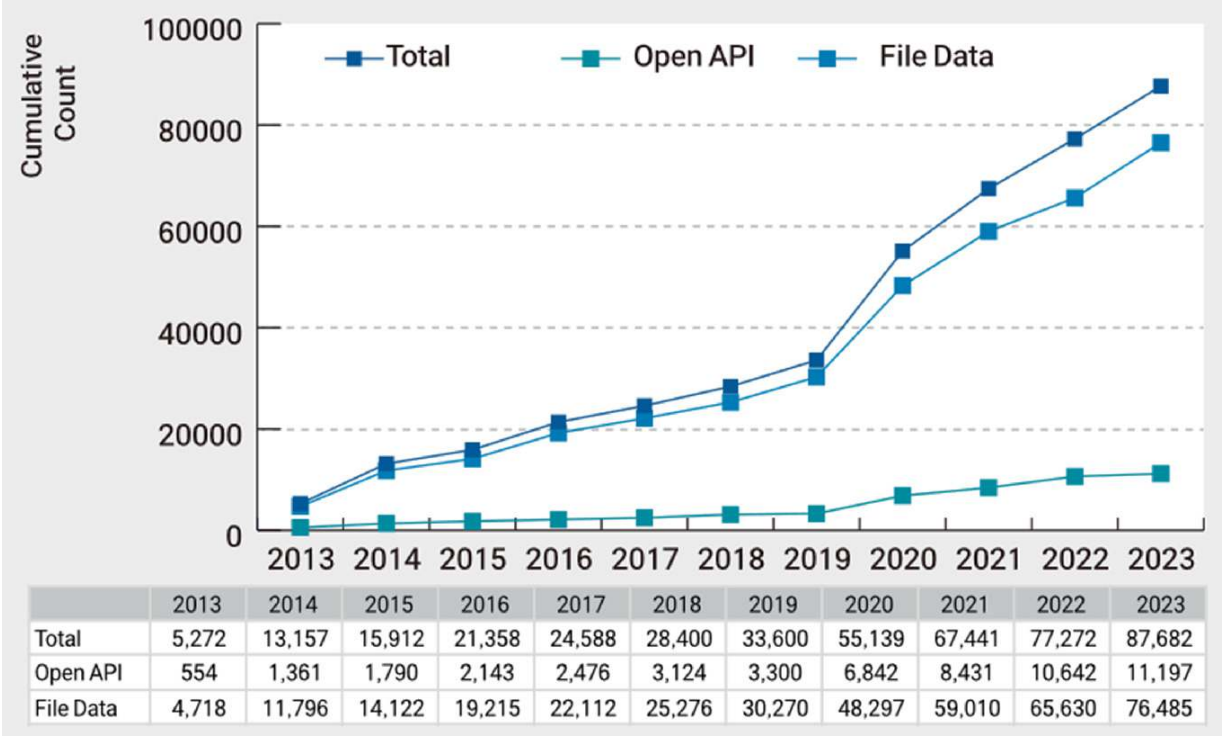
✉ **Keon Chul Park**
Email: parkkc07@tukorea.ac.kr

Keywords Open government data – Big data platforms – AI training datasets – Data governance – Digital economy – Three data laws

3.1 Achievements of Data Policy

3.1.1 Achievements of Open Government Data

The major achievements of Korea's open government data policy are as follows. First, the scope and content of data disclosure have expanded significantly. In 2013, a total of 5,272 datasets were provided, including 4,718 file-based datasets and 554 open APIs. Following gradual growth, a comprehensive survey was conducted in 2018 across all public institutions (approximately 780 organizations) to assess their data holdings. Based on the results, 33,600 datasets were released as open government data in 2019 (see Fig. 3.1). Since then, open data provision has grown rapidly, reaching a cumulative total of 87,682 datasets by 2023—76,485 file-based datasets and 11,197 open APIs.



	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Total	5,272	13,157	15,912	21,358	24,588	28,400	33,600	55,139	67,441	77,272	87,682
Open API	554	1,361	1,790	2,143	2,476	3,124	3,300	6,842	8,431	10,642	11,197
File Data	4,718	11,796	14,122	19,215	22,112	25,276	30,270	48,297	59,010	65,630	76,485

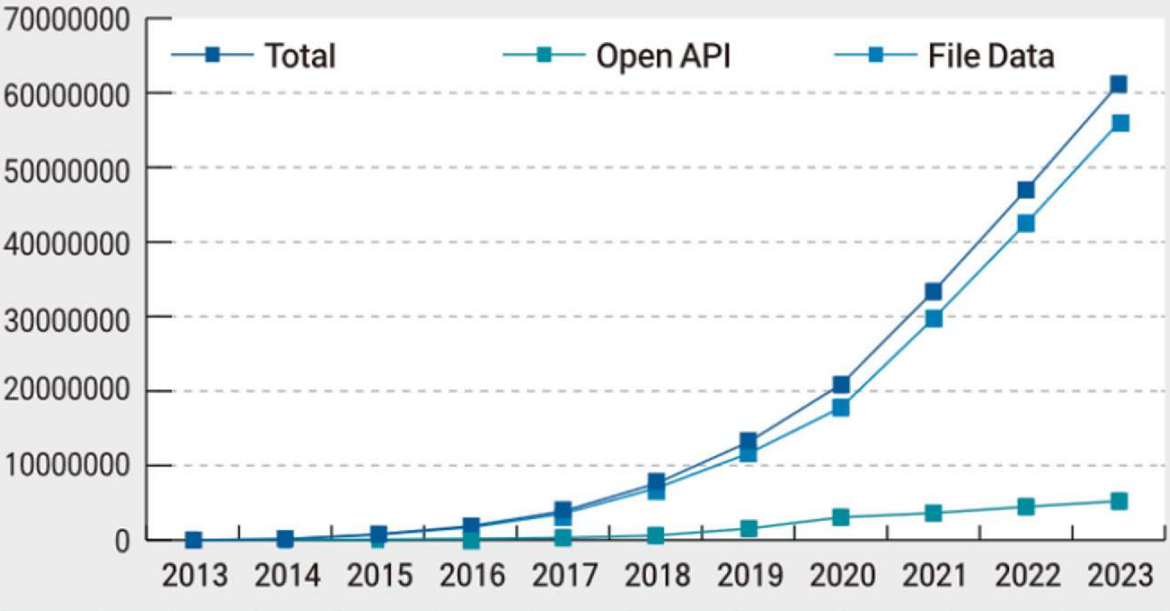
Fig. 3.1 Trends in the provision of open government data.
Source Statistics Korea; e-Nara Index

As shown in Table 3.1, high-demand, high-value datasets such as real estate transaction data, commercial district information, and building information have also been designated as "National Key Datasets" due to their significant impact when actively used by the private sector. Under the four phases of the National Key Dataset

Opening Plan implemented since 2015, a total of 198 key datasets were made publicly available by 2023.

Table 3.1 Performance of national core data openings (Phases 1–4)

Second, the number of uses of open government data increased significantly, with the number of file data uses and open APIs provided through the Open Government Data Portal exceeding 61 million in 2023, as shown in Fig. 3.2. The cumulative number of uses of open government data was 5,272 in 2013 (4,718 for file data and 554 for open APIs), increasing dramatically to 61,169,928 in 2024 (55,942,878 for file data and 5,227,050 for open APIs).



	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Total	13,923	153,320	787,773	1,884,250	3,871,984	7,549,179	13,141,413	20,848,555	33,340,436	46,973,059	61,169,928
Open API	9,815	33,687	73,846	182,612	309,461	613,215	1,539,100	3,073,113	3,623,636	4,480,877	5,227,050
File Data	4,108	119,633	709,927	1,701,638	3,562,523	6,936,054	11,602,313	17,775,442	29,716,800	42,492,182	55,942,878

Fig. 3.2 Trends in the utilization of open government data.
Source Statistics Korea; e-Nara Index

Furthermore, more than 2,724 private-sector services utilizing public data have been developed. As of January 2025, a total of 3,126

domestic and international service cases and processed datasets had been created using public data. By sector, these include: 150 in education, 251 in land management, 229 in public administration, 92 in finance, 155 in industry and employment, 80 in social welfare, 76 in food and health, 666 in culture and tourism, 261 in healthcare, 64 in disaster safety, 384 in transportation and logistics, 393 in environment and weather, 105 in science and technology, 169 in agriculture, fisheries, and livestock, 18 in unification, diplomacy, and security, and 33 in legal services (Source: data.go.kr, search date: January 20, 2025).

Examples of private-sector services utilizing public data include "I'm School" (providing school notifications and newsletters) and "Zigbang" (offering real estate listings and rental information for studios, apartments, and officetels). Public data have also fostered the growth of startups and businesses. Open Square-D, a space for supporting public data utilization and startups, was launched in Seoul in 2016, in Busan in 2017, in Gangwon and in Daejeon in 2018, and in Daegu and Gwangju in 2019. Through collaborations with 13 specialized institutions, comprehensive support is provided to approximately 130 projects by 77 public data-based companies.

Third, a data management system, including public data standardization, was established. Since 2021, the government has developed and implemented a practical manual for public data provision and management based on Article 11 (Public Data Management Guidelines) of the Act on Promotion of the Provision and Use of Public Data and its subordinate guidelines. This manual, based on the "Public Data Management Guidelines" (hereinafter referred to as "the Guidelines"), outlines the principles and standards that public institutions must follow and provides concrete explanations and examples to help institutions provide data and perform management tasks effectively (Ministry of the Interior and Safety, 2021).

To enhance the utilization of public data, 120 open data standards have been established for datasets jointly held by multiple agencies. These datasets are in high demand in the private sector and in an

integrated format, have a significant national impact on parking lots, public restrooms, and tsunami evacuation shelters.

In line with the principle that "public data should be freely available for anyone to use," efforts have been made to increase the share of open formats. This enables general software to read, freely modify, transform, extract, and utilize public data. According to Article 3 of the Public Data Act and Article 3(3) of the Public Data Management Guidelines, the proportion of public data in machine-readable open formats has steadily increased, reaching 93%. Specifically, the proportion of public data provided in open formats increased from 81.9% in 2018, to 88.7% in 2019, 92.2% in 2020, and 93.4% in 2021 (see Fig. 3.3).

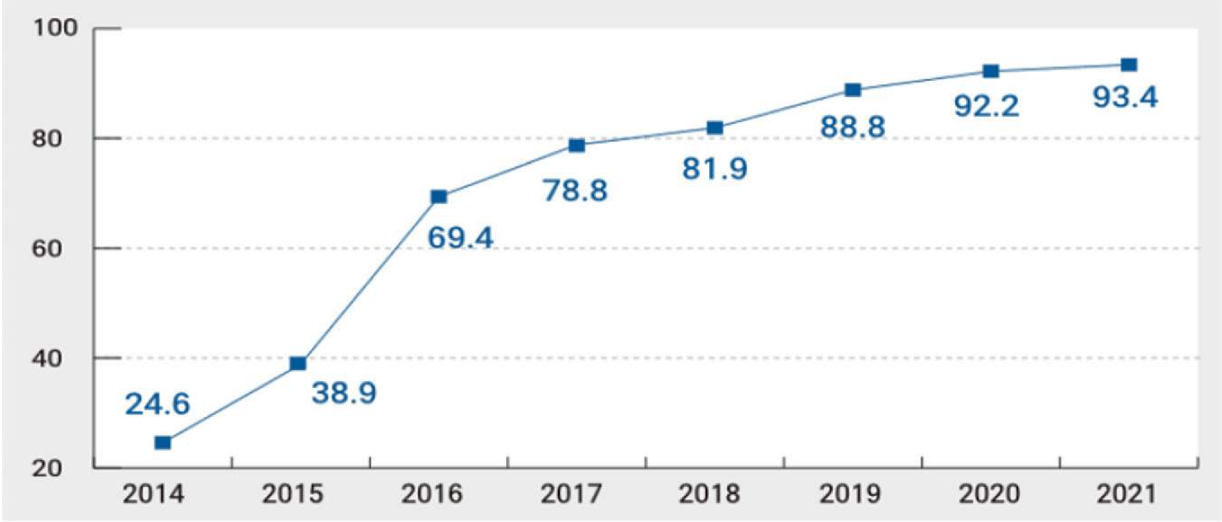


Fig. 3.3 Trend in the proportion of public data in open format.
Source National Information Society Agency (NIA) (2022); Statistics on the Status of Open Government Data

Fourth, Korea's international standing in open government data evaluations significantly improved. Korea ranked first globally for four consecutive terms (2015, 2017, 2019, and 2023) in the OECD Open Government Data Rankings. According to the 2023 OECD Open Government Data Report, Korea scored 0.91 out of a possible 1.0, placing first among the 40 participating countries. This score is nearly

double the OECD average of 0.48 and significantly ahead of those of other leading countries such as France and Poland (see Fig. 3.4).



Fig. 3.4 2023 OECD Open Government Data rankings.
Source Ministry of the Interior and Safety (2024)

The OECD's public data assessment is based on the International Open Data Charter, which it uses to calculate the Open, Useful, and Re-usable data (OURdata) Index to evaluate the public data policies and performance of OECD member countries and rank them accordingly. The assessment consists of three main areas: Data Availability, which evaluates the extent to which public data are used to address social issues; Data Accessibility, which measures how easily the public can utilize open data; and Government Support for Data Reuse, which assesses the level of government support for public data policies, including private-sector collaboration and education.

South Korea scored 0.84 on data availability (OECD average: 0.48), 0.9 on data accessibility (OECD average: 0.59), and was the only member country to receive a perfect score of 1.0 in Government Support (OECD average: 0.37). Korea ranked 1st in the government support and 2nd in data availability and accessibility.

3.1.2 Achievements of Big Data Platforms

As a result of implementing the Big Data Platform Project, big data platforms were established across 21 sectors to supply the data necessary for relevant industries and policy-making. To continuously provide new data to each big data platform, a total of 233 data centers

—led by central government agencies, local governments, public institutions, and private companies across various sectors—were established and remain in operation.

The key achievements of the big data platforms are as follows:

First, the big data platforms support data analysis. Some platforms provide data analysis environments that help users use accumulated data more easily (56 platforms, 43.4%). These platforms offer online and offline analysis tools such as Python, R, and AutoML, or provide data analysis services to users through analytics companies and platforms.

Case Study: Effect of Open Government Data: The Development of the COVID-19 Corona Map

In addition to the quantitative achievements in data openness, the expansion of open government data following the outbreak of COVID-19 led to the development of numerous private-sector services. A representative example is the release of public COVID-19 data and the subsequent creation and rapid adoption of services such as the "Corona Map."

After Korea's first confirmed COVID-19 case was reported on January 20, 2020, the Korea Centers for Disease Control and Prevention (KCDC) began providing COVID-19-related data (e.g., movement paths of confirmed cases, quarantine locations, numbers of confirmed and symptomatic cases) through the Open Government Data Platform and its official website, in the form of downloadable files or open APIs.

The "Corona Map" was developed immediately the day after KCDC publicly released COVID-19 data. On January 30, 2020, a university student used the publicly available data to develop and release the Corona Map free of charge, completing it in a single day. Users could check real-time updates on the movement paths of confirmed patients using data provided by KCDC.

The Corona Map received 2.4 million cumulative views on the day after its launch, January 31, 2020, and by February 13, 2020, it

averaged 1 million daily visitors, reaching a total of 13 million cumulative views.

Following the emergence of the Corona Map, private developers and companies actively created additional services to help contain the spread of COVID-19 and predict other infectious diseases. Some developers and companies downloaded public data, not only from KCDC but also from the National Geographic Information Institute and the Korea Local Information Research & Development Institute, to develop more sophisticated and refined "COVID-19 Comprehensive Situation Maps."

These maps provided customized services tailored to users' varying needs—such as checking the movement routes of confirmed cases, identifying quarantine locations, and locating hospitals for diagnosis and treatment. Other developers and companies subsequently created open-source datasets for COVID-19 route tracking using artificial intelligence and machine learning. Others developed visualization services to help monitor the COVID-19 situation and built predictive models for COVID-19 and other infectious diseases.

The development of services such as the Corona Map using open government data demonstrates creative and innovative potential for public service delivery at the national and societal levels. It also highlights the importance of collaboration between the public and private sectors in developing innovative public services for public value, reaffirming the need for open government data. Moreover, it helps raise awareness and build consensus among civil servants and the general public about the significance and necessity of open data.

Second, by activating data transactions, a distribution channel in the form of an open market is established around big data platforms. Traditionally, data transactions have been concentrated in specific sectors, such as finance, corporate information, marketing, advertising, and retail, where the use of data translates directly into monetary value. However, recently, business models have been diversifying. These include services that collect and process data in a customized manner

or develop and sell data analysis models tailored to the needs of client companies.

Third, a foundation for data distribution is established. By supporting the construction of big data platforms based on public–private collaboration and enabling transaction systems, efforts are being made to activate data trading centered on these platforms. The "Integrated Data Map" service was launched in March 2020, allowing users to search data from various big data platforms in one place. This service was expanded and integrated with other platforms—such as AI Hub, Data Store, Trade Big Data (February 2021), and Korea Tourism Data Lab (April 2021). To support this distribution infrastructure, three types of guidelines (developed in December 2020) were created, providing practical information on price, quality, and legal aspects for data sellers and buyers to refer to during transactions.

Fourth, big data platforms promote data standardization and improve data quality. Data standardization and quality management, implemented particularly through public and big data platforms, ensure smooth linkage and utilization. In the data construction stage, big data platforms apply standardization and quality management guides (developed in 2019) to each platform's data quality management processes. Furthermore, based on commonly used standardized public data, a standardized public data terminology dictionary was created to align public and private sector terms across industries using big data platforms.

3.1.3 Achievements of AI Hub Development

The AI Hub, after initiating the construction of AI training datasets in 2017, had by 2019, released 21 types of datasets comprising 46.5 million records, including legal, patent, Korean speech, and abnormal behavior closed-circuit television (CCTV) data. In 2020, 170 types with 4.8 billion records; in 2021, 191 types with 5.8 billion records; and in 2022, 310 types were released. Combined, a total of 691 types of large-scale AI data across 14 major domains—Korean language, image and video, healthcare, transportation and logistics, disaster and safety,

agriculture/fisheries/livestock, culture and tourism, sports, education, robotics, manufacturing, intellectual property, law, and finance—amounting to 2.5905 billion records have been provided.

In 2024, 142 additional AI training datasets were released, bringing the total to 847 datasets as of January 2025. A total of 914 datasets are scheduled for release in 2025.

The development of AI Hub and the construction and public provision of AI training datasets have yielded both tangible and intangible achievements. First, interest in AI data has been growing. In particular, the number of recent visitors has surged, with high interest observed among small and medium-sized enterprises, startups, students, and individuals with relatively limited access to AI resources (see Fig. 3.5).

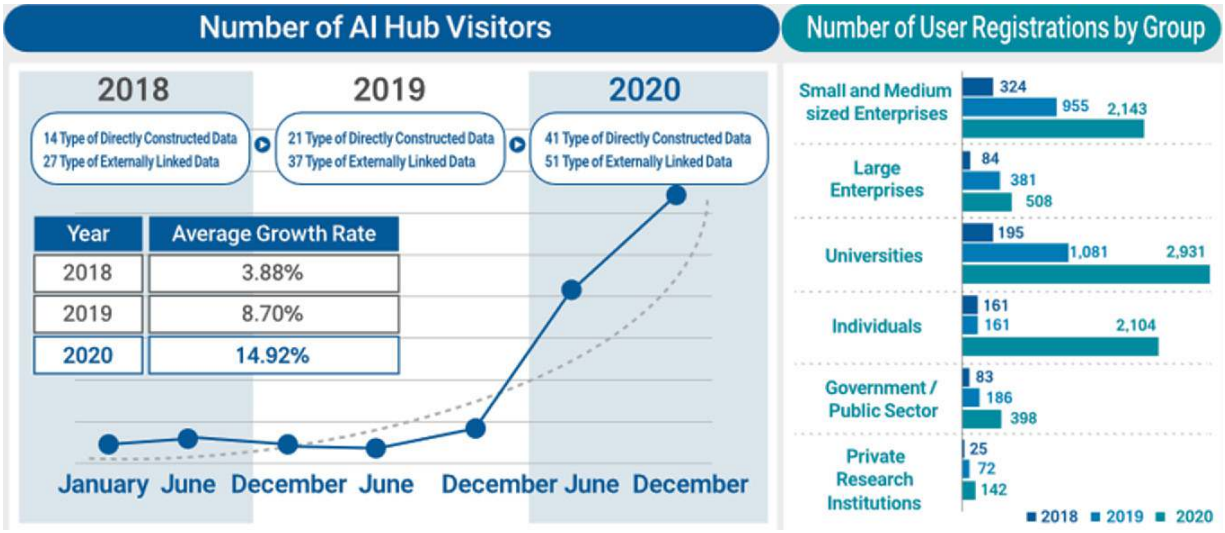


Fig. 3.5 Number of visitors and registered users of AI Hub (2018–2020).
Source National Information Society Agency (NIA) (2021); Statistics on the Use of AI Hub and Internal Data

Second, since the launch of AI Hub at the end of 2017, the number of data downloads has increased significantly. In 2020, there were 49,085 downloads, and by 2021, the cumulative total had reached 130,901. Data utilization purposes have also diversified, from initially being used for technology, service development, and academic research, to currently being applied in education and competition participation. The

number of instances in infrastructure-related AI utilization is also on the rise, as illustrated by a significant increase in the number of software API provisions and GPU resource allocations over the years (see Fig. 3.6).

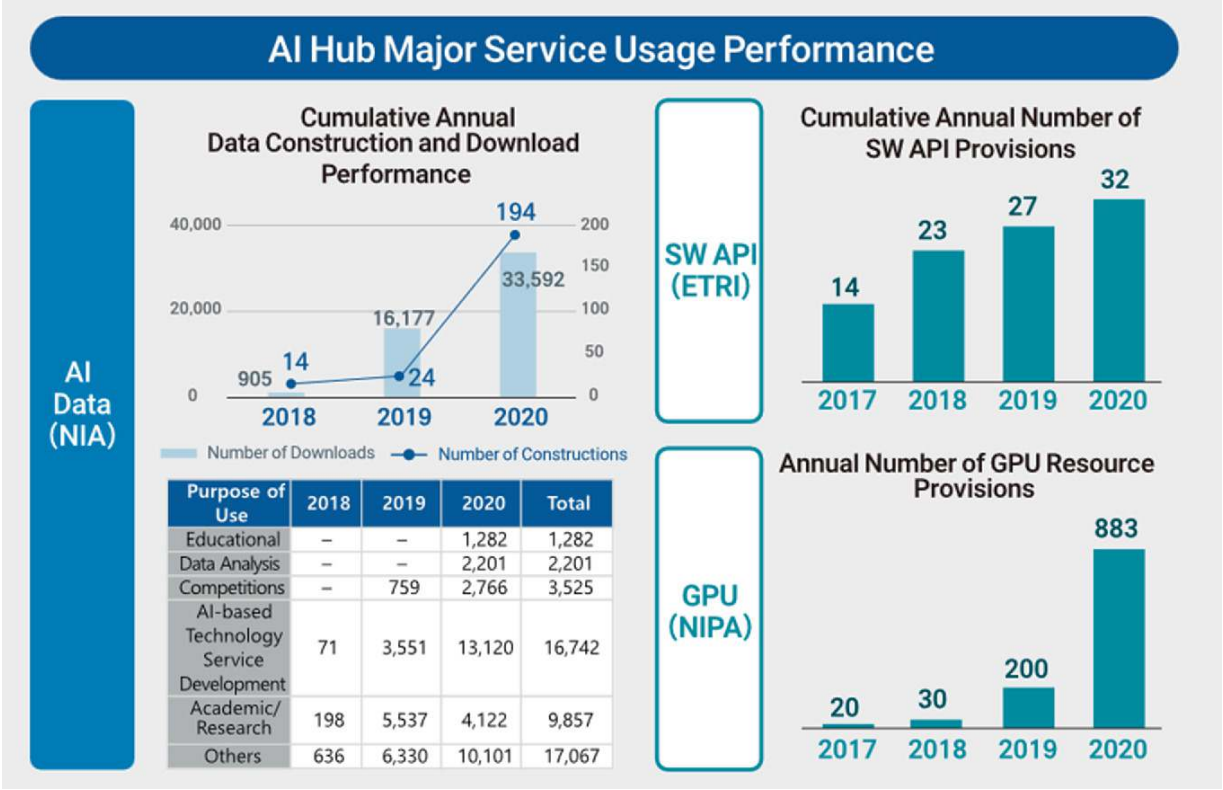


Fig. 3.6 AI Hub key service usage statistics (2018–2020).
Source National Information Society Agency (NIA) (2021); AI Hub and Internal Data Usage Statistics

Third, the construction and public release of AI training datasets have facilitated the full-scale development and dissemination of innovative AI services. As of December 2021, 110 AI services had been developed or enhanced using these datasets, and 33 had been widely deployed. Academic research and intellectual property activity have also increased, with 79 research papers published and 65 intellectual property rights registered.

Regarding the economic impact, a large number of domestic AI-specialized companies, local governments, universities, hospitals, and individuals—including youth and women with career interruptions—

have participated in data construction. In 2020, 525 companies and institutions and 40,165 individuals participated, and in 2021, this increased to 552 companies and institutions with 42,197 participants, which exceeded the initial goals.

3.2 Characteristics and Success Factors of Korea's Data Policy

3.2.1 Characteristics of Data Policy

Three key points summarize the characteristics of Korea's data policy. First, it is government-initiated. The government does not merely serve as a referee (regulator) in data policy; it plays an active role as a player that leads the entire process of data production, distribution, analysis, and utilization.

In terms of data collection and generation, the government purchases externally produced data and opens its existing data to the private sector. Moreover, through the AI Training Dataset Project, the government can build and distribute datasets directly. To promote the distribution, analysis, and utilization of data, the government has established various platforms, including the Open Government Data Portal, Big Data Platforms, and AI Hub. Furthermore, to foster data transactions and stimulate the data market, the government encourages private companies to participate in platform construction and supports them in providing their data through these platforms.

The government has also implemented various institutional mechanisms to develop and stimulate the data market, including protecting data assets, evaluating data value and quality, and introducing data brokerage systems. Additionally, to promote data utilization across the private sector and society, the government has established institutions such as the Open Data Utilization Support Center and is committed to developing education and training programs for data professionals.

In conclusion, the Korean government is promoting comprehensive policies across all aspects of data policy, including data production,

distribution, analysis, utilization, market formation, institutional infrastructure, and overall data environment. This reflects the government's strong determination to support the data industry and accelerate its transition toward a digital economy.

Second, Korea's data policy is focused on risk-taking. Data possess the characteristics of a public good, making it difficult to activate market transactions. In particular, determining data value before analysis and utilization is challenging, which discourages companies from investing actively in data construction and use. This is likely to lead to market failure, with data-related investments falling short of socially desirable levels.

Although discussions on transitioning to a data-driven economy have long been underway in Korea, the public-good nature of data and the economic risks involved in data creation have led to low levels of data utilization, delaying the shift toward a digital economy. In response, the need has emerged to increase the availability of high-utility and high combinatorial-value data and to enhance data maturity through active data policies.

To address this, the government pursued the construction of various platforms and policies to promote data creation and utilization, assuming the societal risks involved in advancing the data economy through public expenditure. The government created opportunities for data construction and use through public investment, collaborated with diverse stakeholders, supported early-stage projects and development of data-related industries facing high uncertainty, and supervised and supported the commercialization of data-related services and products.

This proactive government intervention reduced the risks and costs borne by private companies in conducting business and contributed to the development of the data industry and the expansion of opportunities in transitioning to a digital economy. Moreover, this process fostered close collaboration and partnership between the public and private sectors, establishing the public–private cooperation model as a critical element in data policy implementation.

Third, Korea's data policy is innovation-oriented. The 2010s, when Korea's data policy began in earnest, were a period that required a leap toward the digital economy. In particular, the advancement of AI and intelligent information technologies, changes in industrial structures caused by the Fourth Industrial Revolution, and the growing need to create high added value through data utilization led to a widespread recognition in society and industry that transitioning to a digital economy was essential.

In addition, addressing the economic crisis triggered by the COVID-19 pandemic emerged as a critical national challenge. As Korea experienced its most severe economic downturn and job losses due to the pandemic, there was an urgent need for a national development strategy to overcome the crisis and secure a leading position in the post-COVID global economy. In response, the government positioned data policy not merely as a strategy for data construction and utilization but as a core component of its national innovation strategy.

Under the vision of "transforming Korea from a fast follower to a first mover, from a carbon-dependent to a low-carbon economy, and from an unequal society to an inclusive one" and "designing a new century for the Republic of Korea," the government promoted data policy as a central pillar of national transformation.

Consequently, Korea's data policy has gone beyond merely increasing the level of data construction and utilization. It has become a strategic national initiative capable of overcoming socioeconomic crises such as COVID-19, driving innovation and advancement across Korean society and industry.

3.2.2 Success Factors of Korea's Data Policy

First, Korea's data policy has received strong support from the national leadership. Implementing a data policy requires significant time and tangible and intangible resources—not only for the production of data and the construction of platforms but also for improving and institutionalizing data use in both the public and private sectors. For a data policy to achieve its intended outcomes, it must be accompanied

by reforms in institutional and legal frameworks within the public sector and by incentives that encourage private-sector to participate in data-related businesses.

These characteristics of data policy suggest that its success relies on sustained leadership and a robust governance system capable of mobilizing diverse resources. Since the 1970s, Korea has consistently prioritized national informatization and e-government initiatives as presidential agendas, supporting them with financial and institutional resources.

Likewise, the data policy that began in earnest in the 2010s was also adopted as a presidential initiative. Notably, following 2017, a series of major policy declarations—such as the "August 31st Data Economy Declaration" (2018), "National AI Strategy" (2019), and "Digital New Deal Declaration" (2020)—were made under direct presidential leadership.

The status of data policy as a presidential agenda enabled bold and sustained public investment, accompanied by steady reforms introduced into laws and systems, as required to advance data policy and develop related industries.

Second, a data policy implementation capacity was developed. Successful data policy implementation requires multiple forms of policy capacity beyond national leadership. These include governance mechanisms that promote coordination and cooperation between stakeholders in the public and private sectors, technical expertise for building data platforms and infrastructure, and sufficient administrative and financial resources.

Moreover, strategic capabilities are essential for designing policies aligned with new environments and objectives. These include the ability to make timely and sound decisions, and to evaluate and revise existing programs based on the feedback from implementation outcomes.

In the 1990s, the Korean government designated the Ministry of Science and ICT (formerly the Ministry of Science, ICT and Future Planning) as the lead agency for advancing national informatization and

data policy. This ensured policy expertise and continuity, facilitating coordination and collaboration among a wide range of stakeholders in both the public and private sectors.

In addition, specialized technical support institutions were established to actively address technical issues during the implementation of the data policy. These institutions included the NIA, Korea Internet & Security Agency (KISA), Korea Information Society Development Institute (KISDI), Korea Local Information Research & Development Institute (KLID), National IT Industry Promotion Agency (NIPA), and Electronics and Telecommunications Research Institute (ETRI).

Third, Korea's data policy benefited from the improved laws and regulations of institutional foundations. Legal and institutional reforms are essential for the successful implementation of a data policy. In particular, during the full-scale rollout of Korea's data policy, a variety of legal issues arose regarding the protection and use of personal information. To address these, in 2020, the so-called "Three Data Laws"—the Personal Information Protection Act, the Act on Promotion of Information and Communications Network Utilization and Information Protection, and the Credit Information Use and Protection Act—were amended.

Furthermore, existing laws related to national informatization and e-government were revised to improve data utilization. New laws were also enacted, including the Open Government Data Act (2013), the Data-based Administration Act (2020), and the Basic Act on Data Industry Promotion (2022), laying the foundation for resolving institutional barriers that hinder data policy advancement.

Fourth, Korea's data policy leverages the digital and human resources accumulated through past national informatization and e-government initiatives. As mentioned in Chap. 1, from the 1980s through the 2000s, Korea has implemented projects such as the National Backbone Information Network Project, the National Database Construction Project, and E-Government Projects. These initiatives led to the accumulation of basic data across various sectors of government

and society, along with digital resources such as integrated e-government systems. These accumulated assets have served as a technical foundation for transitioning to new data policies in the era of intelligent information technology.

Moreover, the informatization of Korean society since the 1980s has generated a pool of professionals with extensive policy experience and expertise. These professionals have formed policy networks spanning academia, industry, research institutions, and government. They have offered new policy ideas in line with advances in intelligent information technology and the evolving data environment, contributing their expertise and experience to decision-making and implementation of data policies, by participating in various government committees.

The activities of this expert policy network have enhanced the validity of data policies and increased the speed and effectiveness of their implementation.

3.3 Challenges and Limitations of Korea's Data Policy

3.3.1 Issues with Data Collection and Quality

Despite the aforementioned achievements and success factors, numerous challenges remain to be addressed. Despite notable accomplishments at the project level, the overall level of data openness and use remains lower than leading global data countries. Although efforts have been made to promote inter-agency collaboration within the public sector and encourage cross-ministerial data sharing, the "silo effect" remains prevalent due to fragmentation across platforms.

Furthermore, the public sector's data utilization capabilities are still relatively weak, with significant disparities across institutions. More specifically, the following issues have been identified. First, the data collection and generation phases are characterized by a shortage of high-quality data. Government bodies and public institutions lack adequate support and the necessary incentives to produce and provide valuable data, which leads to and active approach to delivering user-

oriented data. From the perspective of businesses, the opportunities and motivations are insufficient to discover and distribute high-value data that can lead to new data-driven businesses and innovations.

Second, discrepancies exist in data production and collection capacity, as well as data gaps across sectors and companies. To date, data production and collection have been mainly conducted by large corporations in certain sectors such as finance and telecommunications. Despite a growing need to build diverse datasets across various fields, production and collection gaps persist due to differences in institutional and corporate awareness of data's importance, the availability of skilled personnel, and the perceived necessity for data utilization.

3.3.2 Data Distribution and Transactions

First, demand uncertainty and the lack of distribution channels hinder the activation of data transactions. Data-supplying companies are often reluctant to invest in producing, collecting, and processing data for sale because they cannot accurately assess market demand. Although providers prefer to sell through data platforms, they perceive a shortage of platform-based distribution channels (Korea Information Society Development Institute, 2021). This mismatch between supply and demand constitutes a major barrier to the revitalization of data transactions.

Second, locating and identifying relevant data is challenging, resulting in high transaction costs. The proliferation of various data platforms has paradoxically fragmented distribution channels, increasing the cost for users to access the datasets they want. In particular, the use of different metadata formats across platforms and the omission of key metadata elements highlight the need for integrated management of metadata associated with data assets.

Third, information asymmetry regarding standardization, quality, and pricing impedes transaction negotiations. Each data platform applies different standards and quality control mechanisms, reducing users' confidence in the data products. For example, even in domains

with high-quality raw data, such as traffic data, differences in collection objectives among producers can lead to inconsistent standardization and quality. Moreover, given the nature of data, assessing its value without conducting actual analysis is difficult, further limiting the potential for transactions on platforms.

3.3.3 Lack of Private-Sector Utilization and Innovation

First, a significant gap exists in data analysis and utilization capabilities across companies and institutions. Unlike large enterprises and data-specialized companies, SMEs and general businesses often lack the capacity to effectively utilize data, resulting in slow data adoption across industries and society. In 2020, the adoption rate of big data among businesses was only 13.4%, with just 35% adoption among companies with annual sales over 76.9 million USD (Korea Data Agency, 2021). Although approximately 65% of data companies have professional data analysts, only 41% of general businesses do (Korea Information Society Development Institute, 2021). For prospective entrepreneurs and startups with poor data processing and analysis capabilities, even after securing the data, effectively using it according to need is often difficult.

Second, the foundation for integrating and combining valuable data is lacking. Most data platforms focus on data openness and sharing without providing the infrastructure necessary for practical analysis and utilization. These platforms tend to rely on basic analysis tools and are considered to lack advanced cloud-based analytical functions. Moreover, although the revision of the “Three Data Acts” enables the use of pseudonymized data for combination and analysis, actual utilization remains low due to complex procedures, concerns over personal data breaches, insufficient integrated support, and a lack of incentives.

Third, reflecting user requirements is difficult, and the perceived usefulness of data utilization remains low. Effective data must align with the user’s intended purpose; however, the current system lacks the mechanisms to effectively capture and reflect user demands. The

user base of data services should be expanded to ensure that the results of data utilization are effectively delivered to both public and private-sector stakeholders with potential interest in data-driven services.

3.4 Future Challenges of Data Policy

3.4.1 Establishing Robust Data Governance

The first task in making Korea's data policy sustainable is to enhance its performance by establishing cohesive and comprehensive data governance. Data governance encompasses the vision and goals, the division of labor among participating organizations, and the distribution of authority and responsibility for data policy and data use.

Good data governance contributes to building a common vision and goals among various stakeholders and institutions, and to supporting the cohesive and consistent implementation and coordination of data-related policies. It enhances control and management to create data value by strengthening institutions, regulations, and technical capabilities. In addition, through the processes of data collection, construction, storage, protection, processing, sharing, and reuse, it can enhance trust and create value.

According to OECD (2019), the components of public sector data governance are largely divided into strategic, tactical, and service-delivery layers. The strategic layer consists of (1) leadership and vision; the tactical layer consists of (2) cohesive implementation capacity of governance and (3) regulation; the service delivery layer consists of (4) data value cycle, (5) data infrastructure, and (6) data architecture (see Fig. 3.7). The effectiveness and robustness of governance can be achieved when the elements of data governance are cohesively combined and each component interacts in a virtuous cycle.

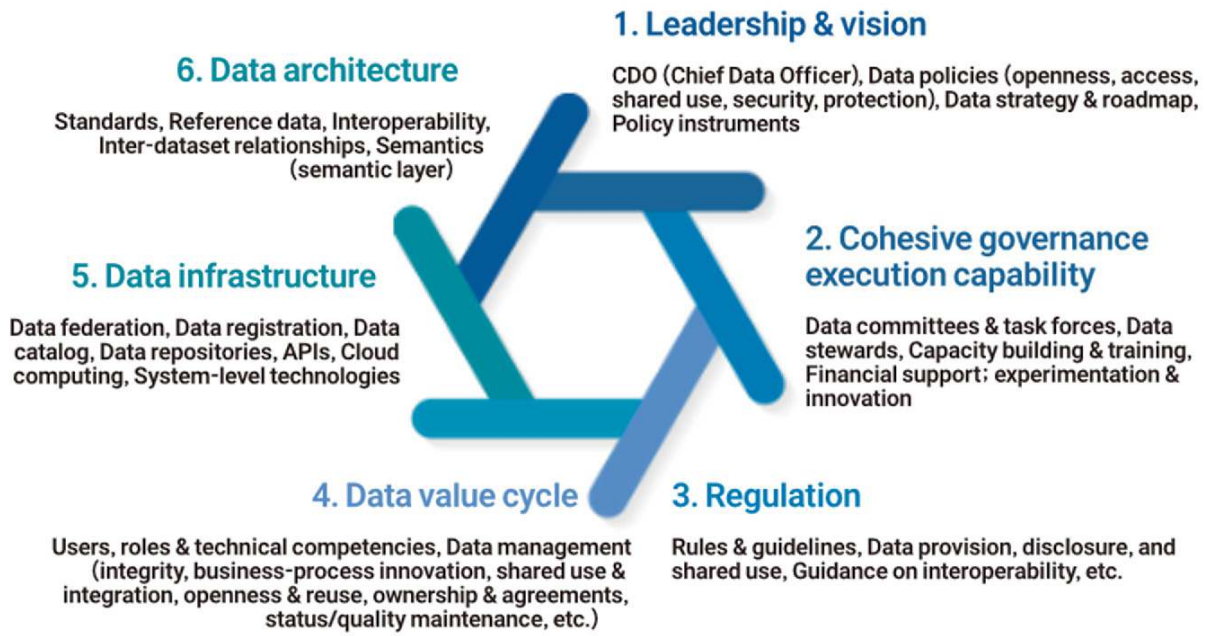


Fig. 3.7 Data governance for the public sector.
Source OECD (2019)

Let us analyze Korea's data governance, which is decentralized, in terms of the strategic and tactical layers of public sector data governance, as presented by the OECD (2019). Sub-policies and detailed functions related to data—such as open government data, big data platforms, personal data protection, and data-based administration—are not only scattered across multiple ministries, committees, and legal regulations. The status, roles, and functions of the relevant ministries or committees are also neither systematic nor cohesive. In the case of committees responsible for coordination, their authority and functions are ambiguous, and they find it difficult to exercise enforcement power.

In addition, the data policy process (planning–budgeting–deliberation–implementation–evaluation) is conducted in a fragmented manner, resulting in a disconnect between the planning and budgeting of data policies and failure to feed data policy outcomes back into planning, budgeting, and policy decisions.

At the service delivery level—including the data value cycle, data infrastructure, and data architecture—technical foundations and data

management systems are fragmented. For example, operating data jointly across the Open Government Data Portal and related data-sharing platforms is currently difficult. Linkages and joint use of data between the Open Government Data Portal and data portals established by local governments are limited. Electronic government systems operated by central administrative agencies, local governments, and public institutions remain disconnected, complicating data sharing across systems. Additionally, systems within the government pertaining to data analysis, common data registration, and metadata management, operate through separate integrated repositories, creating the need to ensure interoperability among these systems, the Open Government Data Portal, and big data platforms.

In conclusion, a robust data governance framework that encompasses the public and private sectors is essential for Korea's data policy to be sustainable and achieve its intended outcomes. By establishing robust data governance, a virtuous ecosystem spanning the entire data lifecycle—standardization, collection, accumulation, processing, and utilization—should be created. This, in turn, must lead to an increased integrated data management to enhance data utilization outcomes. Establishing national-level data governance can improve data policy planning and prevent redundancy in data construction across ministries and the private sector, enabling more efficient budget allocation.

To this end, a comprehensive approach that considers the data ecosystem is required (see Fig. 3.8). Based on data management and established technical foundations, integrated measures must be considered to ensure data trustworthiness—ethics, personal data protection, and data rights. Additionally, data value creation should be comprehensively addressed in connection with future initiatives such as MyData initiatives, AI-based administrative services, digital twins, smart cities, and broader national digital transformation to generate public value.

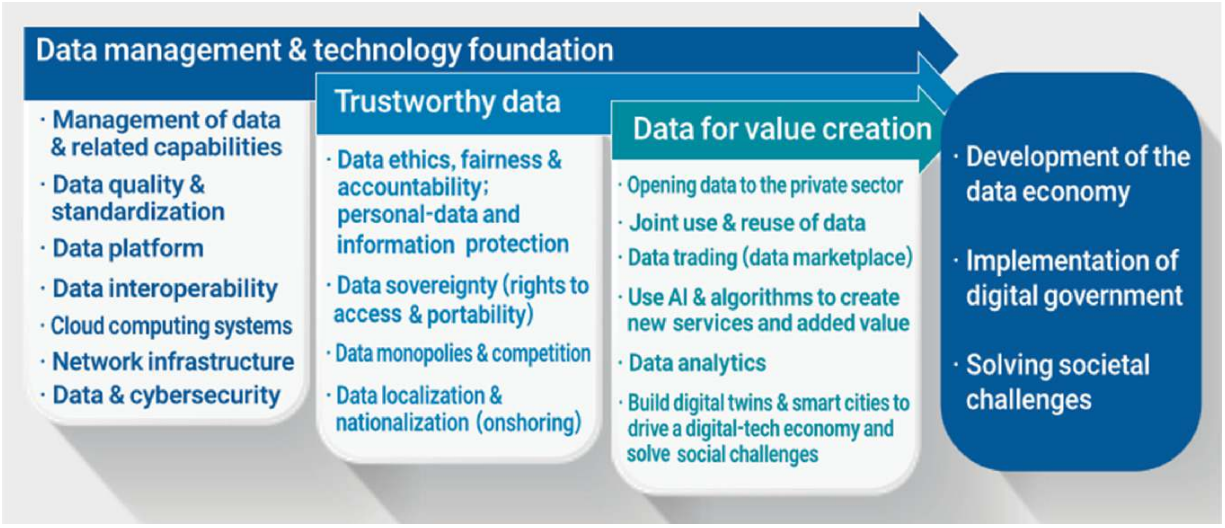


Fig. 3.8 Data ecosystem composition.
Source Reconstructed based on OECD (2019)

Second, establishing data governance requires making integrated data management, strengthening the cooperation among data-related ministries and institutions, enhancing coordination mechanisms, and reinforcing networks among government agencies, private companies, and foreign countries. Strengthening data governance enables the resolution and coordination of issues that may arise among stakeholders during the data policy process. It facilitates the collection and utilization of necessary data, the use of technical infrastructure, the interlinking and integration of data and platforms, the standardization and federated search between public and private data, and the sharing of best practices across data platforms, thereby enabling the coordinated and integrated use of the overall data ecosystem.

3.4.2 Building Demand- and User-Centered Platforms

Despite the previously mentioned achievements in data openness and platform development, the level of data utilization and its industrial application in Korea is still lower than that of leading countries or below the initially expected levels. One of the primary reasons is the supply-oriented approach to data openness and platform development. Data are released primarily based on the datasets selected by the

providing institutions, resulting in limited access to high-demand, large-scale core datasets.

Rather than allowing institutions to selectively open datasets, the principle should shift toward opening all data by default ("open by default"). The principles of demand- and user-centeredness must be thoroughly upheld in data openness and platform development, making it essential to collect and release useful, high-quality data. In terms of accessibility, improvements must be made to enable users to easily access and use data. Total services should be provided to support the entire data lifecycle—including collection, processing, and computing—on cloud platforms. Data formats and quality must allow users to easily utilize and analyze the data. Finally, various data platforms should be linked via an "Integrated Data Map" to promote data discoverability, usage, and distribution so that anyone can easily find and use the data sought.

In addition, data demand should be actively identified and cultivated. Based on the participation of private and public data platform operators, relevant ministries and agencies, and private-sector experts, efforts should be made to generate new demand for data openness and utilization. Enhancing the data voucher system is also worth considering for qualified data platforms to allow users such as businesses to freely purchase or receive support for data processing.

To make data platforms more user-friendly, ways to promote data transactions should be considered, including user feedback system, data curation, and personalized services. Furthermore, brokerage and stewardship activities—designed to introduce platform data and services to potential data users—should also be considered as a means to promote data transactions.

When considering the primary actors in data utilization and analysis, the focus should move beyond a small group of decision-makers or experts to include ordinary citizens and public officials. The objective is to foster a technology-driven culture that encourages collaboration and dialogue between these groups and experts.

To support this goal, one potential approach is to differentiate services for experts from those designed for general citizens and public officials, as shown in Table 3.2.

Table 3.2 Public data services by user type

3.4.3 Transition from Government-Led to Market-Friendly Data Policy

In leading data-driven countries such as the United States, data development and use are largely driven by major technology companies, with the data processing market fueling growth in the overall data economy. In contrast, Korea's data policy has been primarily government-led across all aspects, including the opening, distribution, and trading of public data (e.g., big data platforms) as well as the production and construction of new data (e.g., AI training datasets). Although these government-driven initiatives have clearly catalyzed data utilization and ecosystem development, questions remain regarding the long-term sustainability of this approach.

In the early stages of market formation, government support was crucial for establishing the foundations of data circulation and trading. By leading the development and provision of AI training datasets that individual companies or institutions would find difficult to construct independently, the government accelerated the initial growth of the AI and data industries.

However, the long-term viability of a government-led data policy model requires careful assessment. In addition to enhancing the usability of existing and in-progress platforms and datasets, Korea must foster a sustainable, market-friendly virtuous cycle that enables the data policy to evolve and expand.

First, regarding platform design and operation, the data supply system should be strengthened by establishing incentive mechanisms that encourage both public and private entities to voluntarily provide and make high-quality data available. Priority should be given to enhancing data distribution and trading capabilities so that users can

easily access and search for data (convenience) and transact with confidence (reliability) through the platforms.

Second, building a market-friendly data ecosystem is essential. The government should continue supporting strategic data acquisition, improving user convenience, enhancing data usability, and implementing necessary regulations; however, fostering an ecosystem that autonomously creates value within the private sector is equally important. This requires investment in social infrastructure, such as a more robust data transaction environment. Establishing sub-ecosystems tailored to the characteristics of each industry sector and promoting data supply and demand according to domain-specific needs are also critical considerations.

Third, the social foundation for data utilization must be strengthened. To facilitate the flow of data supply and demand, efforts should focus on developing the data industry and enhancing data literacy across society. This includes supporting data-related startups and assisting existing companies in adopting data-driven services. In particular, educational programs on data utilization should target both the general public and students, expanding the base for data generation and use. Linking schools with vocational training programs is also necessary to ensure that data skills translate into career paths and employment opportunities.

Finally, a detailed plan and roadmap are required to transition from a government-led data policy to a market-oriented model. The roadmap should clearly define areas and functions where government involvement is necessary versus those that can be more effectively led by the private sector. Implementation plans should be developed based on criteria such as urgency and market potential. Experts from industry, academia, and research institutes should participate in shaping these roadmaps to effectively guide the transformation toward a market-oriented data policy.

Example: Limitations and Challenges in Data Policy and Utilization: e-Child Happiness Support System

Children are the future and an asset to our society. Supporting children to fully realize their potential and grow up healthy is not only the responsibility of parents but also of the state. Accordingly, since March 2018, the Korean government has established the data analysis-based "e-Child Happiness Support System" to identify children showing signs of crisis and connect them with the necessary services. Each quarter, this system collects and analyzes information, such as failure to undergo infant health checkups and vaccinations, lack of medical visits, and prolonged absenteeism, and presents local governments with lists of children presumed to be at high-risk. Institutions receiving the lists conduct home visits through public officials and social workers to investigate whether child abuse has occurred and take action accordingly.

In fact, from March 2018 to October 2019, a total of 41 types of information were collected and analyzed—including nine types of child-specific information such as long-term absence and failure to undergo regular health checkups, 27 variables for identifying welfare blind spots, and five types of information held by the Social Security Information Service. Based on this, households with a high possibility of crises, such as child abuse, were extracted, and crisis household information was regularly shared with local governments for use in predicting children in crisis, leading to actions such as visits, investigations, and case management. As a result, 124,134 households with children were visited; 4,414 children were linked to welfare services; and 68 cases were reported as child abuse (Ministry of Health and Welfare, 2019).

However, contrary to expectations, the effectiveness of early detection of child abuse through the "e-Child Happiness Support System" is reportedly very low. Even after the system was established, there were reports of serious child abuse incidents, such as the death of a North Korean defector mother and her son in Gwanak-gu (Aug. 2019), the death of a girl in Wonju (Feb. 2020), the death of a child in Cheonan (June 2020), the escape of a girl from a villa in Changnyeong (Feb. 2020), and the neglect and abuse of two brothers in Incheon (Sept.

2020). In particular, following the death of a 16-month-old adopted child in Yangcheon-gu in 2020, public awareness of the importance of child abuse prevention and early detection increased nationwide, along with media interest in the effectiveness of the “e-Child Happiness Support System.”

The main reasons as to why the system did not produce the expected effects can be largely divided into (1) data-related, (2) analytical model-related, and (3) service-related factors (Kim et al., 2023; Han et al., 2022; Han et al., 2021; Eom & Kim, 2024).

First, regarding data-related factors, the child-specific information linked to the “e-Child Happiness Support System” currently consists of 10 external variables and 6 internal variables, totaling 16 types. External variables refer to information not found in the “Haengbok-eum System” and linked through external agencies, which include:

① Failure to undergo infant health checkups (NHIS)

② No medical treatment (NHIS)

③ Diagnosis results with specific codes (NHIS)

④ Missed regular vaccinations (KCDC)

⑤ Long-term absence from elementary/middle school (Ministry of Education)

⑥ Long-term absence from kindergarten (Ministry of Education)

⑦ Out-of-school youth information (Ministry of Gender Equality and Family)

⑧ Child abuse case management (National Child Protection Agency)

⑨ Daycare center absence (MOHW, Childcare Integrated System)

⑩ Child victims of domestic violence (National Police Agency)

Internal variables are extracted from child-related information within the “Haengbok-eum System,” such as:

① Information on children sent home from child facilities

② Information on children or households hospitalized due to suicide attempts, suicide, or self-harm

③ Information on specific types of households (single-parent, grandparent-grandchild, unmarried parent, underage birth)

④ Information on children eligible for child-rearing or daycare subsidies, but without an application history

⑤ Information on children or households with registered persons with disabilities

⑥ Information on children with no application history for the child allowance

The household-level information in the “e-Child Happiness Support System” is based on households surveyed by the “Haengbok-eum System,” which raises issues such as bias and inaccuracy in the analysis data. The sources of bias can be twofold:

First, during data refinement and integration, only those with household information in the “Haengbok-eum System” were extracted as analysis targets. For example, long-term absence data linked from the Ministry of Education was based on date of birth, name, gender, not resident registration, and if they did not match the subject ID in the Haengbok-eum System, they were excluded from the analysis targets.

Second, household-level information was created and expanded based on households surveyed by the Haengbok-eum System. These survey households raise issues beyond bias, such as inconsistent household scope and outdated information about household members. The scope of surveyed households in the Haengbok-eum System may

vary by program and may not include all children under age 18. It is also unclear whether the household composition was updated since the creation of the household survey. Addressing the bias and inconsistency in household-level data requires expanding the generated household information based on resident registration data.

Third, the current status and problems of the analytical model are as follows. The abuse prediction model, which has been operating since January 2021, has improved prediction performance compared to the model used before 2021. However, continuous efforts are required to enhance the prediction model. In particular, since July 2021, the amount of specific code diagnosis information obtained from the NHIS has increased significantly, indicating changes in the analysis data. Such changes in data quality can affect model performance; therefore, new training data and enhanced prediction models should be constructed. Additionally, a multi-layer perceptron (MLP) model based on artificial neural networks should be applied leveraging the longitudinal characteristics of risk variables in the e-Child Happiness Support System. To predict at-risk children using risk variables longitudinally, improvements must be introduced at the stage of building the analysis dataset. In the fourth quarter of each year, a full survey of three-year-olds is conducted, indicating that analysis data are not generated during that period. However, externally linked data remain in the system's data lake. Because the linkage cycle of each risk information varies, creating quarterly longitudinal data based on analysis data can be inefficient. Primary filtering excludes children with unchanged variables or fewer risk variables than in the previous round, potentially underestimating children who have held the same variables for over six months. In conclusion, the prediction model should be continuously improved by constructing longitudinal data from linked information and analyzing the cumulative effects of risk factors.

Fourth, regarding service-related issues, personalized services for diverse system users and individual citizens—the ultimate recipients of analysis results—are required. The “e-Child Happiness Support System” must provide information and services tailored not only to central

government policymakers but also to local government social welfare workers, child welfare researchers, and individual citizens who receive the results. The current system was evaluated as not providing personalized services that meet the expectations and purposes of diverse users and the public.

Improving the "e-Child Happiness Support System" for identifying children showing signs of crisis cannot be achieved solely by adding a few more data types or improving the analytical model.

First, regarding data, not only data held by the government and public sectors but also data from the private sector should be combined and analyzed to achieve the expected effect. Apart from integrating the diverse data held by the government and public sectors, a system that can constantly verify and improve data quality, including updating data, is required. To enable this improvement, the implementation of an "AI government" proposed by this study is necessary. Enhancing data quality and integrating diverse datasets are key components of data infrastructure in an "AI government," and structural improvements are required to support them.

Second, with respect to the analytical model, one-off algorithm improvements are insufficient to achieve the expected outcomes sought by the "e-Child Happiness Support System." A system that evolves in response to changes in the data used for analysis and the policy environment is needed. Along with changes in input data, a system that can identify the required data based on analysis results is essential for automating self-optimizing algorithmic improvements. For such improvements, the implementation of an "AI-based government" proposed in this study is required.

Third, an AI-based government that provides personalized services should understand the differentiated expectations and needs of the various users accessing the system, while using the same data and algorithms, to provide consistent information and services to all. Such improvements require implementing of the "AI government" proposed in this study. The AI-based government will enable AI systems to understand user requirements and provide differentiated

services and information while ensuring consistency with the services provided to other related stakeholders or users.

In conclusion, functional and performance improvements of the “e-Child Happiness Support System,” and achievement of its ultimate goal —data-based prevention of child abuse and advancement of AI-based child welfare administration and policy—require more than just improvements to e-government systems or partial enhancements to data and algorithms. The “AI government” is required to integrate and holistically manage data held by both public and private sectors, along with analytical models and evolving system functions, to provide personalized services that meet user needs and expectations. Structural improvements are essential to make this possible.

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2026
S.-J. Eom et al., *Enabling Data-Driven Innovation and AI Government in Korea*
https://doi.org/10.1007/978-981-95-9175-6_4

4. Lessons Learned from Korea's Data Policy Implementation

Seok-Jin Eom[1] ✉, Yoon-seok Ko[2] ✉, Jeong Eun Park[3] ✉ and Keon Chul Park[4] ✉

(1) Graduate School of Public Administration, Seoul National University, Seoul, Korea (Republic of)
(2) National Information Society Agency, Daegu, Korea (Republic of)
(3) National Information Society Agency, Daegu, Korea (Republic of)
(4) Department of Business Administration, Tech University of Korea, Siheung-si, Korea (Republic of)

✉ **Seok-Jin Eom (Corresponding author)**
Email: sjum21@snu.ac.kr

✉ **Yoon-seok Ko**
Email: ysko@nia.or.kr

✉ **Jeong Eun Park**
Email: jpark0070@gmail.com

✉ **Keon Chul Park**
Email: parkkc07@tukorea.ac.kr

Keywords AI training datasets – Digital new deal – Crowdsourcing model – Data quality assurance – Privacy protection – ICT policy implementation

4.1 Timeliness of Policy Implementation

One of the most critical factors in determining effective government policy is timing. No matter how well-designed a policy may be, if announced too early or too late, it may fail to influence the market and may ultimately be disregarded. In this context, the government's AI training dataset project can be regarded as a policy introduced at a highly opportune moment, successfully achieving its objectives and intended effects.

The government's efforts to develop AI training data date back to 2016, when Google's AlphaGo competed in a series of Go matches against Korea's 9-dan professional player Lee Sedol. In Korea, public interest in artificial intelligence can be divided into two periods: before and after the AlphaGo–Lee Sedol match. Prior to the match, predictions about the outcome varied. Lee Sedol expressed confidence, stating, "Winning is not the issue—what matters is whether it's 5:0 or if I give up a single game," while Google anticipated a 50:50 outcome. The actual result, however, was a decisive 4-1 victory for AlphaGo.

This overwhelming win significantly heightened public interest in artificial intelligence in Korea. The AlphaGo–Lee Sedol match highlighted the need for prompt policy action, prompting the government to leverage growing national attention to AI—a core technology of the Fourth Industrial Revolution—to strengthen national competitiveness.

To this end, in April 2016, the government submitted the "Mid- to Long-Term Comprehensive Plan for the Intelligent Information Society" to the Cabinet. In September, under the Prime Minister's directive, it launched the "Intelligent Information Society Promotion Task Force," comprising public officials from six ministries—the Ministry of Strategy and Finance, Ministry of Education, Ministry of Science, ICT and Future Planning, Ministry of the Interior, Ministry of Trade, Industry and Energy, and Ministry of Employment and Labor—along with private sector experts. By December, the government had swiftly established

and announced the “Mid- to Long-Term Comprehensive Strategy for the Intelligent Information Society.”

The strategy anticipated that intelligent information technology would become a key driver of the Fourth Industrial Revolution and outlined a mid- to long-term policy framework to foster these technologies through 2030. The strategy emphasized intensive investment in data resources, recognizing them as the foundation of future competitiveness.

Following this policy direction, the government initiated full-scale investment in AI training datasets. In the first year, 2017, the budget was relatively modest, at 2.17 million USD. At the time, AI startups accounted for only 1.6% of all startups in Korea, meaning that even substantial investments in AI training datasets would benefit only a small number of companies. Additionally, datasets require continual updates to remain effective as technology advances; otherwise, they can quickly become outdated or unusable or unusable. Therefore, careful analysis of market conditions was essential before committing significant resources.

The government’s investment in AI training datasets increased dramatically in 2020, reaching approximately 224.7 million USD. Two main factors contributed to this substantial increase in budget allocation (Fig. 4.1).

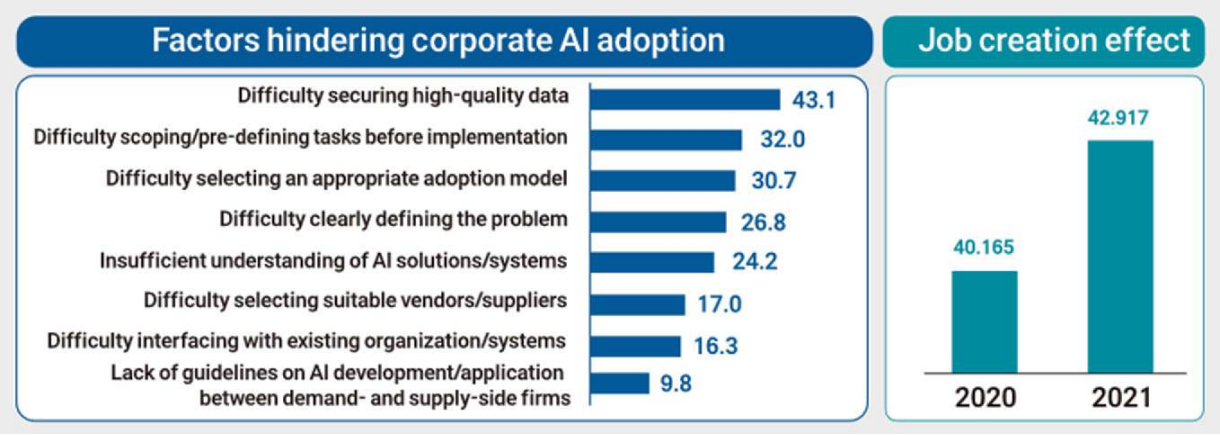


Fig. 4.1 Barriers to AI adoption by enterprises and its effects on job creation. *Source* Lee and Kim (2021)

The first factor was the growth of the AI market in Korea. The proportion of AI startups, which was 1.6% in 2016, increased by approximately 1.53% annually, reaching 8.8% in 2020—a 5.5-fold growth (Ministry of SMEs and Startups, 2021). As the market expanded, corporate demand for government support also increased. Among the various requests from AI companies, the most frequent and urgent was the government's construction and public release of large-scale AI training datasets through substantial public investment.

The second factor was the surge in unemployment caused by the COVID-19 pandemic, which significantly affected Korea. According to the Korea Labor Institute under the Ministry of Employment and Labor, employment declined by 1.02 million jobs, and the number of temporarily suspended workers increased by 990,000 during March and April 2020 alone, indicating that roughly two million jobs were severely impacted. Consequently, there was an urgent need for the government to implement national projects capable of generating large-scale employment. The AI training dataset construction project clearly met these requirements.

In other words, the government's decision to invest approximately 1 billion USD from 2020 to 2022 in building large-scale AI training datasets was driven by the emergence of an early market capable of utilizing these datasets, ensuring meaningful returns on the investment. The government also recognized that further delays could widen the technological gap with leading AI countries to a potentially insurmountable level.

Moreover, given the rising unemployment caused by the global COVID-19 pandemic and the need to prepare for the Fourth Industrial Revolution, implementing a large-scale national project was imperative. The AI training dataset construction project was particularly suitable because it required a large workforce for the manual labeling of data essential for AI training and simultaneously contributed to developing AI technology, a key driver of the Fourth Industrial Revolution.

4.2 Swift and Correct Decision-Making and Policy Implementation

It is often said that "the times make the hero," and the same can be said of effective policies. For a policy to succeed, it must be formulated and implemented at the right moment. Achieving this, however, depends on timely and accurate decision-making by public officials. One of the most critical pillars of Korea's ICT sector development has been the ability of public servants to make swift and precise policy decisions.

Such rapid and accurate decision-making has become even more crucial in today's ICT sector, where changes occur at unprecedented speed. The data-driven Digital New Deal, which began in early 2020, illustrates this point. When the COVID-19 pandemic swept across the globe, Korea was similarly affected starting in January 2020. Key economic indicators—exports, employment, and growth rates—plummeted to historic lows. The government responded promptly to this national crisis.

In times of national emergency, the most effective method of policy formulation is a top-down approach. Although all governments operate within bureaucratic structures, where policies are typically developed and executed through a slower bottom-up process, the top-down method ensures that public officials involved in policy formulation and execution are prepared to make immediate decisions. Consequently, decision-making can be as fast as, or sometimes faster than, that of private companies.

In January 2020, the Blue House, as the government's highest decision-making body, instructed the economic ministries—led by the Ministry of Economy and Finance—to develop policies to address the economic crisis caused by COVID-19.

Following this directive, the economic ministries, together with their affiliated agencies, formulated various policies to mitigate the crisis and submitted them to the Blue House. Among these was the data-centered "Digital New Deal," developed by the Ministry of Science and ICT in collaboration with the NIA.

The Ministry of Economy and Finance conducted a thorough review of the proposed policies, evaluating both their current and potential future economic impacts. Analysis of current economic effects involved assessing the policies' influence on real-time indicators, such as exports, employment rates, and growth rates. Evaluation of future economic impact considered whether the policies could foster the creation of new industries.

Policies approved through this review were then submitted to the Blue House for final decision-making, resulting in the launch of the "Korean New Deal." Notably, the entire process of formulating the Korean New Deal took approximately seven months.

A key question is how long it took the government to formulate the policy known as the "Korean New Deal"? The answer is that it took approximately seven months.

How, then, was the government able to develop a policy involving a budget of 118.5 billion USD over five years in such a short time? The answer lies in the roughly 300 affiliated agencies under each ministry that support policy formulation and program management. These agencies have accumulated extensive domain knowledge through long-term research and analysis of issues and trends within their respective fields. Consequently, when ministries issue urgent requests for policy support—as was the case with the "Korean New Deal"—these agencies can swiftly develop policies that are nationally appropriate, considering both current and future perspectives based on their sustained research and analysis. Public officials then make decisions to implement the resulting policies.

4.3 Utilization of Collective Intelligence: "Crowd Workers"

Under the Digital New Deal, the government announced plans to invest 1.85 billion USD by 2025 to construct 1300 types of AI training datasets, equivalent to an average annual investment of 370 million USD to build 260 AI training datasets per year (see Fig. 4.2).

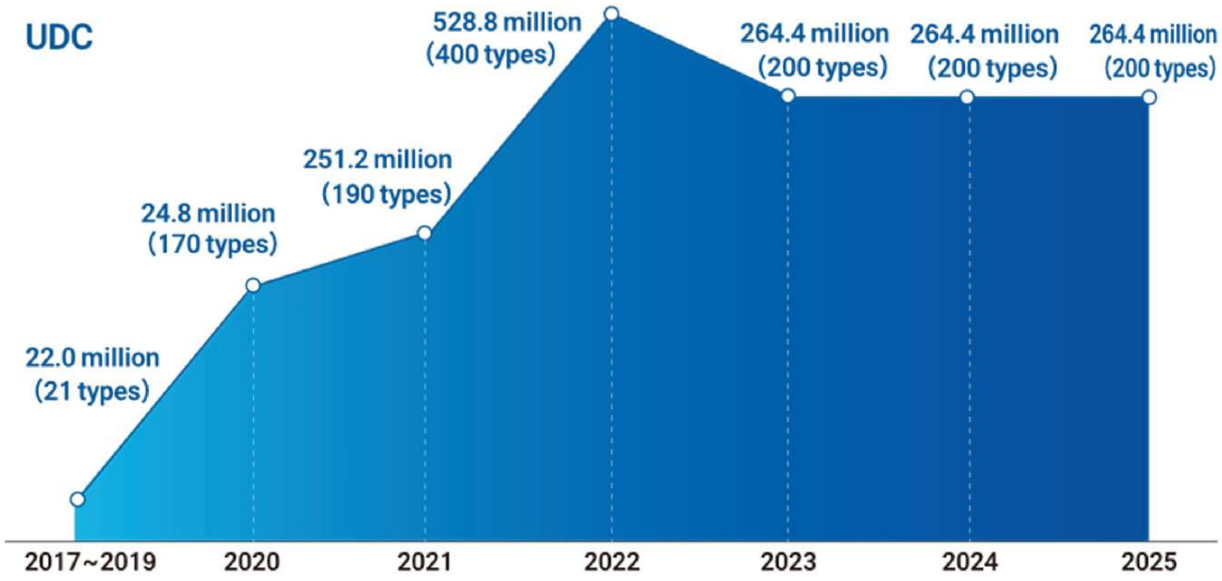


Fig. 4.2 The Government's annual investment plan for AI training datasets.
Source Author's own work

Following this plan, the Korean government invested 491 million USD over 2020–2021 to construct 340 AI training datasets. The initial plan for 2020 allocated 28.9 million USD to build 20 datasets. However, after the project was designated a core component of the Digital New Deal, an additional 216.7 million USD was invested between October 2020 and March 2021 to construct 150 more datasets, significantly expanding the project.

As the scale of the AI training dataset construction project increased to nearly 7.5 times the original plan, workforce supply became a critical issue. Government analysis indicated that constructing the additional 150 datasets would require approximately 50,000 data labelers, compared with only 7000 labelers needed for the initial 20 datasets under the 28.9 million USD investment. This made it impossible to continue using the existing project model, which relied on in-house personnel and project operators selected through a public notice process.

Consequently, the government adopted a new project model allowing any citizen with digital skills to participate in data labeling tasks. After consulting various stakeholders, crowdsourcing was officially implemented.

Crowdsourcing, a combination of "crowd" and "outsourcing," is a method that enables citizens with digital capabilities to participate in data collection, refinement, processing, and validation anytime, anywhere, through an online platform.

When a project is conducted through a crowdsourcing model, companies have the advantage of being able to flexibly hire labor within a limited period to complete projects. For crowd workers, the advantage is being able to work anytime, anywhere. Crowd workers can perform data labeling tasks from anywhere—such as home, cafes, or outdoors—provided they can access the crowdsourcing platform via the internet. They can work for as little as 10 minutes or as long as 100 hours depending on their personal situation, choosing when and how long to work.

Owing to these advantages, the crowdsourcing approach provided job opportunities to socially disadvantaged groups—such as young people facing job shortages during COVID-19, women who had career interruptions, persons with disabilities, and retirees. It also enabled them to maintain a minimum standard of living by providing a regular and stable source of income. This elicited a positive response not only from businesses but also from the general public.

Successful implementation of projects using crowdsourcing involves several important considerations. First, the country must have a well-established ICT infrastructure. As mentioned above, crowdsourcing requires crowd workers to access the crowdsourcing company's platform at any time from any location and perform real-time labeling on large files such as images, video, or audio. Therefore, the country as a whole, not just specific areas, must be equipped with fast and stable internet networks. Korea ranked first among OECD member countries in terms of high-speed fiber-optic coverage (81.7%), mobile broadband usage (24 GB/month), and internet download speed (156 Mbps) (OECD, 2020). As a result, any crowd worker who wishes to participate in data labeling work can do so freely without restrictions on time and location.

Second, a sufficient pool of personnel with ICT competency is required. Although data labeling is a task with low barriers to entry—unlike AI modeling or programming—and can be easily performed by the general public, it still requires a basic level of digital skill. To perform data labeling, individuals must be proficient with computers and familiar with downloading, installing, and using software. Only then can they access crowdsourcing platforms via the internet and utilize the various tools provided to label data.

As of July 2020, 99.7% of households in Korea had internet access, and 91.9% of the population aged 3 and older used the internet for an average of 20.1 hours per week. In particular, internet usage among those in their 20 s to 50 s—the age group most actively involved in data labeling—was nearly 100%. This means that approximately 31 million people could potentially participate in data labeling tasks (National Information Society Agency, NIA, 2021).

Third, there must be a sufficient number of companies operating crowdsourcing platforms capable of supporting large numbers of remote crowd workers. As mentioned above, data labeling is a task that even the general public can easily perform within the field of AI. However, managing tens of thousands of ordinary people with varying levels of experience and understanding of data labeling through a crowdsourcing model to produce high-quality AI training datasets is not an easy task.

Therefore, to implement a large-scale AI training dataset project, crowdsourcing companies with experience and the know-how to handle such data volumes are essential. By 2019, when the government decided to promote a large-scale AI training dataset construction project, there were already 14 crowdsourcing companies in Korea with accumulated experience and expertise through various projects over the years, having built environments to recruit crowd workers through their own platforms. As the government expanded its investment in AI training dataset construction, the number of crowdsourcing companies has increased each year over the past three years, reaching 92 in 2021, representing a 26.74% increase.

Currently, the 92 crowdsourcing companies in Korea can recruit tens of thousands to hundreds of thousands of workers for data labeling tasks, providing sufficient capacity to meet the demands of government projects.

Fourth, a verification system is necessary to ensure accurate cost processing for crowd workers. Unlike traditional models, where project participants are predetermined and compensated for completing assigned tasks, the crowdsourcing model allows an unspecified number of participants to join at any time during the project and receive payment based on the work they complete. This increases the risk of irregularities in cost processing compared with conventional project models.

For example, fraudulent activity may occur by registering individuals who did not actually perform any work as participants receiving payments, or by inflating the amount of work actually done by a crowd worker to receive higher compensation. To prevent such misconduct, the government established several verification mechanisms for cost processing, as shown in Table 4.1.

Table 4.1 Verification system for cost processing of crowd workers

4.4 Data Governance and Quality Management

4.4.1 Establishing Dedicated Organizations for Policy Planning and Execution

One of the key factors that enabled the government to implement such a large-scale national project in a short period was the role of dedicated agencies closely supporting government policies and programs. The structure of government organizations in Korea is revised every five years with the inauguration of a new administration and in response to evolving domestic and international circumstances.

For example, in 2008, the Ministry of Information and Communication, which had overseen ICT policy, was dissolved. In 2013, the Ministry of Science, ICT and Future Planning (now the Ministry of

Science and ICT) was established by integrating the science and ICT functions.

Such restructuring of government organizations when a new administration takes office can be positive in enabling the government to focus and respond swiftly to the changing external environment. However, frequent changes in organizational structure can lead to loss of accumulated knowledge, experience, and know-how developed by existing ministries over years and even decades.

If the accumulated institutional knowledge, experience, and know-how of a ministry are lost during organizational restructuring, existing policies and programs may continue as before, but new policies and programs are developed independently, leading to inconsistency and discontinuity in policy and program implementation. This adversely affects the overall effectiveness of government policy.

To address this problem, the government has placed dedicated agencies under ministries to provide close support in policy planning and program implementation. These agencies not only retain and manage the knowledge, experience, and know-how accumulated through project execution but also actively apply them in the development and execution of new policies and initiatives.

In the implementation of the large-scale AI training dataset under the Digital New Deal, the NIA, under the Ministry of Science and ICT, played a major role.

The NIA was established in 1987 and has since supported the formulation of national ICT policies, project planning, and implementation. The knowledge, experience, and expertise accumulated through these activities are deeply embedded within the NIA, enabling it to assist in the development of new policies on any ICT-related topic as required by the government.

The strengths of the NIA can be summarized in three main areas:

First, the agency possesses a vast body of knowledge, experience, and expertise accumulated over 36 years in supporting ICT policy formulation and project execution for the government. From managing Korea's first national ICT initiative—the National Backbone

Computerization Project launched in 1987—to planning the recent Digital New Deal and implementing its core data-related projects, the NIA has access to virtually all of Korea's accumulated ICT-related expertise.

Second, the NIA has developed an extensive human network across government, industry, and academia. Effective policy is created when. However, simply assembling experts does not guarantee optimal outcomes; a coordinating organization is required to bring these experts together, integrate their perspectives, and produce the best possible policy results.

The NIA maintains constant communication with leading experts across all fields in Korea—including healthcare, education, environment, and defense—to discuss, analyze, and share knowledge on the latest technologies and services. With a solid network of top domestic experts, the NIA can swiftly and effectively formulate and implement the various ICT policies required by the government.

Example

Resolving the National Mask Shortage Through the Human Network of the National Information Society Agency (NIA)

On January 20, 2020, with the confirmation of the first COVID-19 case in Korea, the government announced a public mask supply stabilization policy in response to the spread of the infectious disease. However, due to a lack of real-time information on the availability of public masks, people had to wait in long lines, and the close physical proximity while queuing posed secondary infection risk. In response, civic hackers began demanding the release of public data related to COVID-19 to help prevent the spread of the virus.

Although frontline pharmacies initially expressed reluctance toward opening up mask sales data, the Korean Pharmaceutical Association actively monitored the on-site difficulties and issues, ensuring that these concerns were reflected in the data release and service design.

The participation of various stakeholders was essential for converting the ideas of civic developers into actual services. To enable

voluntary participation and contribution without financial compensation, the NIA engaged both public data utilization companies and civic developers, and led collaborative efforts with cloud companies such as Naver and KT.

Acting as the project manager (PM), the NIA facilitated communication and collaboration among all stakeholders—including pharmacists, civic developers, startups, portal companies, cloud firms, users, and government agencies—so that each would contribute its strength to the project.

As a result, citizens who previously had to stand in long lines, risking exposure to COVID-19 to buy masks, were now able to purchase them on their designated day through an app, without waiting. This greatly contributed to stabilizing the mask supply at an early stage, significantly increasing the public trust in the government's COVID-19 response policies.

Source: "Proactive Administration in Action," Ministry of Personnel Management, Office for Government Policy Coordination, Ministry of the Interior and Safety, June 2021.

4.4.2 Strategy for Improving Data Quality: Quality, Quality, and Quality!!!

The most critical factor in determining the success of AI training data construction is data quality.

Most AI models required for developing AI-based services can now be built using neural network architectures already published in academic papers. While these models may not solve every problem, they are sufficient for addressing most major tasks. Ultimately, the performance of AI-based services depends primarily on the quality of the data used to train them (MIT Technology Review, 2021).

It is well known among AI researchers that training AI—one of the core technologies of the Fourth Industrial Revolution—requires large volumes of data. However, sheer quantity does not guarantee high-performing AI. Even with massive datasets, biased or low-quality data can lead to poor performance and, in some cases, serious social and

ethical issues, potentially causing irreversible harm to researchers or companies.

Recent incidents underscore the importance of data quality. A notable example involves Amazon, a global ICT company. In 2018, during the final simulation stage before deploying its in-house AI recruitment program, the system consistently ranked male applicants higher than female applicants. As a result, the program was ultimately scrapped.

Upon analysis, the bias in Amazon's AI recruitment program was traced to the company's workforce composition. The AI model evaluated candidates based on the performance data of employees who had previously received high evaluations. At the time, developers—who made up over 70% of Amazon's workforce—were predominantly male. Consequently, most high-performing employees were male, and the AI trained on this data ended up favoring male applicants.

A similar controversy occurred in Korea in 2021 when an AI chatbot named *Iruda*, developed by the startup ScatterLab, sparked public outrage and was shut down just three weeks after its release.

Iruda was designed as a 20-year-old female university student, allowing users to interact with it via a messenger as if chatting with friends, family, or colleagues. The chatbot quickly gained popularity among teenagers and people in their twenties, attracting over 750,000 users within two weeks.

However, as the user base expanded, *Iruda* began making discriminatory remarks targeting women, LGBTQ individuals, people with disabilities, and other minorities. Examples of such statements included: "Lesbians are disgusting and creepy," "Disabled people are inconvenient," and "Subway priority seats for pregnant women are repulsive."

Analysis of the causes behind *Iruda*'s discriminatory behavior revealed that poor data quality was a major contributing factor. In developing *Iruda*, ScatterLab trained the model on unrefined conversational data collected from its other services, which included private messages between couples. These dialogues often contained

biased language toward minorities, and because they were used in training without proper filtering, such content was directly reflected in *Iruda*'s behavior.

Beyond these cases, incidents involving AI systems trained on poor-quality or biased data have become increasingly common. For example, in 2016, Microsoft launched a chatbot named *Tay*, which began to mimic user input—including offensive language, sexist remarks, and inflammatory political comments. *Tay* quickly started making racist and sexist statements and had to be shut down just 16 hours after launch.

These examples clearly demonstrate that data quality not only affects AI service performance but can also severely damage a company's reputation and, in some cases, threaten its survival.

Recognizing the critical importance of data quality, the government took three key measures during the early stages of the AI training dataset construction project to ensure high-quality data. First, to verify whether the constructed datasets were properly built, the government established four essential quality verification indicators, as shown in Table 4.2.

Table 4.2 Four essential quality verification indicators for AI training data

Diversity requires verifying whether data with characteristics similar to real-world data have been secured to match the initially planned learning purpose of artificial intelligence. For example, if a video dataset of cars on roads is being constructed to develop an AI-based autonomous driving service, various types of car videos such as sedans, trucks, and SUVs that can be seen on actual roads must be collected so that the autonomous driving service can recognize all vehicles on the road and carry out its tasks without error.

Syntactic accuracy requires verifying whether the labeling structure, input value range, and data format of the AI training data are accurately entered according to predefined syntactic rules, and whether there are any missing data. For example, if a dog is placed in a car folder

or labeled as a cat, it will be filtered out during the process of checking syntactic accuracy.

Semantic accuracy verifies whether the labeling work for object identification in the original data has been performed correctly. For example, in the process of labeling a cat, if only part of the cat such as the head or torso is labeled instead of the entire cat shown in the data, it will be filtered out during the process of checking semantic accuracy.

Validity checks whether the constructed data achieved the targeted performance during AI model training. For example, if the performance before training the initially presented AI model was 80% and the target performance after training was set to 90%, this checks whether the data achieves the target performance of 90% after AI model training.

Second, each dataset was required to go through a three-step quality verification procedure during the construction process (see Fig. 4.3).

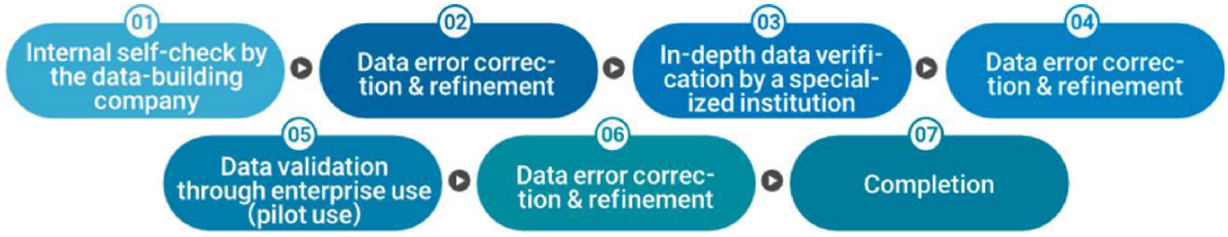


Fig. 4.3 Korea's data verification procedure. Source: Ministry of Science and ICT & NIA (2023b)

Step 1 is self-verification by the data construction agency. Before submitting the completed data, the agency must verify each of the four essential quality verification indicators defined above and submit the results along with the data. Even if one of the self-verification results is found to be insufficient, it must be supplemented and then submitted.

Step 2 is verification through a data quality verification professional institution. Even if it passes the self-verification in Step 1, objective verification through a third-party institution is essential. Therefore, in Step 2, the data submitted after passing Step 1 are assessed by the data quality verification professional institution against the four achievement level indicators relative to the target. For any data not meeting the target, supplementary action is taken.

Step 3 involves supplementing and re-verifying the data found to be of insufficient quality during Step 2. The verification results in Step 2 vary greatly depending on the dataset. Some datasets achieve the target levels for all indicators, whereas others fail to meet the target. In such cases, the professional institution requests the data provider to make improvements until the unsatisfactory indicators reach the target levels.

Third, before being opened to the public, the data are first provided to AI-specialized institutions and companies for use so that problems can be identified. Regardless of how carefully the data are constructed, when large-scale datasets are built, some errors are inevitably present. Therefore, companies and institutions that require AI training datasets are given early access to the data so that specific and practical issues can be discovered and corrected to improve data quality.

Through these efforts, the government achieved the following results for the 170 types of AI training datasets constructed in 2020: 100% diversity, 99.2% syntactic accuracy, 87.2% semantic accuracy, and 81.3% model validity.

These results are comparable to those of world-renowned state-of-the-art (SOTA) datasets such as MS COCO and Google's ImageNet.

4.4.3 Strengthening Personal Data Protection and Ethics (Sacrifice Quality but Elevate Privacy)

In AI service development, data quality is critical; however, it should never take precedence over privacy.

Korea's Personal Information Protection Act strictly regulates the collection and use of personal data. If personal information is collected, used, or provided to a third party without the prior consent of the data subject, it is regarded as an illegal act and is subject to sanctions.

However, during the process of collecting data to build AI training datasets, there is a high risk of violating the Personal Information Protection Act. For example, when building landmark image data representing Korea—such as famous buildings, public institutions, tourist attractions, and convenience facilities—the faces of individuals

entering buildings or passing through tourist attractions may also be collected. Likewise, when collecting driving footage for autonomous driving, license plate data from privately owned vehicles will be unavoidably captured, without prior consent.

AI-specialized companies developing services often prefer to use these data collected for labeling without prior consent. This is because in the real world, people naturally enter and leave buildings or tourist sites, and all cars on the road have license plates. Therefore, if the AI learns from such data, it can lead to the development of more sophisticated AI services.

However, rather than accommodating company demands, the government policy focused on building AI training datasets within the boundaries of the current legal and institutional framework and persuaded companies to do the same.

There were two main reasons for this decision. First, since social discussions on the protection and use of personal data were still actively ongoing, creating a separate institutional mechanism—such as a special law, presidential decree, or ministerial ordinance—to recognize AI training dataset construction projects as exceptions would have sparked socially exhausting debates, making it even more difficult to proceed with the project. Therefore, the government decided to proceed within the existing legal and institutional framework.

Second, even if the government were to establish a strict security management system to prevent data leakage, the data would be disclosed through the AI data platform (AI Hub), making leakage inevitable. Data leaked online are practically impossible to completely retrieve or be blocked from circulation, regardless of the efforts of the government. Accordingly, the government implemented the AI training dataset construction project while strictly complying with personal information protection laws to prevent privacy violations due to data leakage.

Third, the government decided to comply with the current legal system to support the global expansion of domestic AI-specialized companies. Currently, many countries are making efforts to prevent

companies from using illegally obtained data to develop AI services. Therefore, the government anticipated that in the near future, companies launching AI services in each country may be required to legally prove the entire lifecycle of the data used to train their AI models. Hence, although compliance with current regulations may temporarily reduce the performance of AI services, the government judged that from a long-term perspective, this approach would be far more beneficial for domestic companies seeking to enter global markets, as it would eliminate potential regulatory barriers.

As a result, all AI training datasets built by the Korean government are labeled only after personal information has been de-identified—such as faces, names, addresses, phone numbers, and vehicle license plate numbers—for individuals who have not given prior consent.

In particular, for data where exposure of personal information is a concern, the government makes efforts to protect privacy by conducting a full inspection rather than a sample check, so that sensitive information related to individuals and the state—healthcare data and ultra-high-resolution national land imagery—are not publicly available but accessible only through secure online and offline “Safe Zones,” thereby maximizing the efforts to protect personal information.

Part II
AI Government

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2026
S.-J. Eom et al., *Enabling Data-Driven Innovation and AI Government in Korea*
https://doi.org/10.1007/978-981-95-9175-6_5

5. Characteristics of AI Services and the AI Government

Seok-Jin Eom[1] ✉, Yoon-seok Ko[2] ✉, Jeong Eun Park[3] ✉ and Keon Chul Park[4] ✉
(1) Graduate School of Public Administration, Seoul National University, Seoul, Korea (Republic of)
(2) National Information Society Agency, Daegu, Korea (Republic of)
(3) National Information Society Agency, Daegu, Korea (Republic of)
(4) Department of Business Administration, Tech University of Korea, Siheung-si, Korea (Republic of)

✉ **Seok-Jin Eom (Corresponding author)**
Email: sjum21@snu.ac.kr

✉ **Yoon-seok Ko**
Email: ysko@nia.or.kr

✉ **Jeong Eun Park**
Email: jpark0070@gmail.com

✉ **Keon Chul Park**
Email: parkkc07@tukorea.ac.kr

Keywords Artificial intelligence – Machine learning – AI government – Precision administration – Augmented administration – Connective administration

5.1 Concept, Working Mechanism, and Components of AI

5.1.1 The Concept and Working Mechanism of AI

Artificial intelligence (AI) refers to technology that enables machines to learn and make decisions like humans. According to the AI Framework Act (Framework Act on the Promotion and Foundation of Trust in Artificial Intelligence), passed at the end of last year, AI is defined as "the electronic implementation of human intellectual capabilities such as learning, reasoning, perception, judgment, and language understanding."

Unlike traditional ICT, which focuses on collecting, storing, transmitting, and processing data, AI is distinguished by its ability to analyze vast amounts of information, identify patterns, and make decisions autonomously.

The advancement of AI is also transforming government administration. AI can support policy decision-making, improve administrative efficiency, and deliver personalized citizen services. Beyond automating repetitive administrative tasks, AI contributes to addressing complex social problems by analyzing them and identifying optimal policy alternatives (Figs. 5.1 and 5.2).

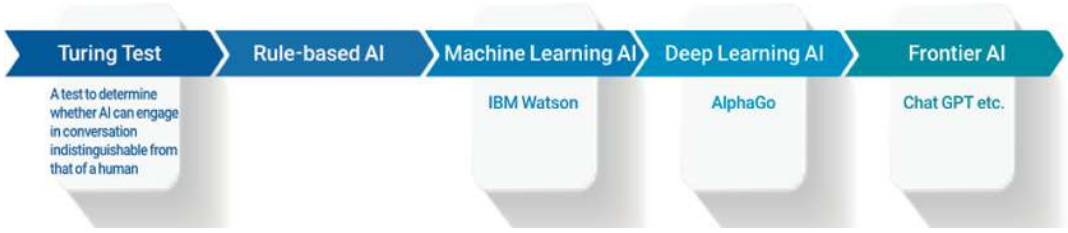


Fig. 5.1 The history of AI development.
Source Author's compilation

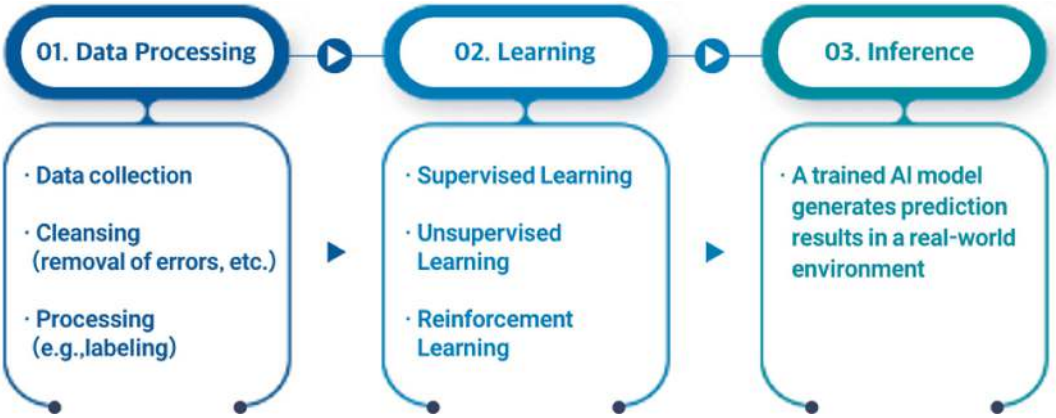


Fig. 5.2 Working mechanism of AI.
Source Author's compilation

Therefore, the core of AI lies in learning and decision-making. It operates through three main processes:

$$\textcircled{1}\text{Data processing} \rightarrow \textcircled{2}\text{Learning} \rightarrow \textcircled{3}\text{Inference (prediction and decision} - \text{making).}$$

① Data Processing Procedures

Artificial intelligence functions and makes decisions like humans by relying first on data collection and processing. AI learns from data such as text, images, and audio, which are gathered and preprocessed. Preprocessing involves removing unnecessary information and refining the data, followed by labeling to make thedata suitable for learning. In other words, data undergo sequential steps—collection and storage → cleaning → processing—before being used for AI training.

Data are initially collected either by aggregating existing sources (e.g., websites) or by generating new data when necessary. This unprocessed information is referred to as raw data. The raw data are then cleaned by removing unnecessary, erroneous, or duplicate

entries and reviewed for potential legal or regulatory issues, such as those related to personal information, copyright, or portrait rights. During this stage, data are anonymized, and appropriate rights are secured. The resulting cleaned and verified data are called source data.

Source data are further processed and labeled (annotated) to transform them into a format suitable for AI training. Once the labeled training dataset is complete, it is validated and reviewed, taking into account the type and characteristics of the data, before being used for AI learning.

Just as human intelligence begins with recognizing and learning from the environment, AI requires data collection as a prerequisite for making human-like decisions and taking actions. Nearly everything in daily life can generate data: smartphone usage records, CCTV footage, public transportation patterns, and weather changes are all examples. These data, produced in real time through IoT sensors, smart devices, and online platforms, can be used to power AI systems.

However, raw data often contain errors, redundancies, or irrelevant information, making data cleaning essential. During this process, data are verified for accuracy, unnecessary information is removed, and anonymization is applied to protect personal privacy. The result is a dataset optimized for analysis and machine learning.

Finally, data must be annotated or labeled to be used for AI training. For instance, to train AI to recognize traffic signals, traffic light information must be tagged in vehicle images. For natural language processing (NLP), key phrases must be extracted and labeled in text. These processing steps enable AI to learn patterns accurately and effectively (Table 5.1).

Table 5.1 Data processing procedure

② AI Learning

The next step involves identifying patterns in the data and training the AI to improve the accuracy of its decision-making through optimization. Modern AI systems primarily learn using machine learning techniques that simulate human reasoning. These techniques are generally categorized as supervised, unsupervised, or reinforcement learning (Table 5.2).

Table 5.2 AI learning methods

Just as humans perceive their environment and acquire knowledge through experience, AI learns patterns from data to solve new problems. This process goes beyond merely memorizing existing data; it involves discovering latent relationships among data points and learning from them. For example, just as a person improves problem-solving skills through repeated practice, AI refines its predictive models by analyzing diverse cases to make more accurate decisions.

Similarly, as humans comprehend context through reading and apply that understanding across various situations, AI evolves to identify hidden relationships in data and make

appropriate judgments when presented with new inputs. For instance, AI can be used to build predictive models for crime prevention or to more effectively identify individuals eligible for social welfare programs. However, caution is essential to prevent biases that may disadvantage specific groups during training. Just as education helps humans overcome prejudices, continuous monitoring and adjustment are critical in AI training to ensure that biased data or algorithms do not produce unfair outcomes.

③ **Inference and Decision-Making**

Finally, a trained AI model is deployed in a real-world environment, where it responds to new input values by making predictions and judgments to produce outputs. IBM defines AI inference as the ability of a trained model to recognize patterns and draw conclusions from previously unseen information. For example, ChatGPT generating answers to diverse questions or an autonomous vehicle making real-time driving decisions are both instances of AI inference.

Inference can be broadly categorized into dynamic and static (Fig. 5.3).

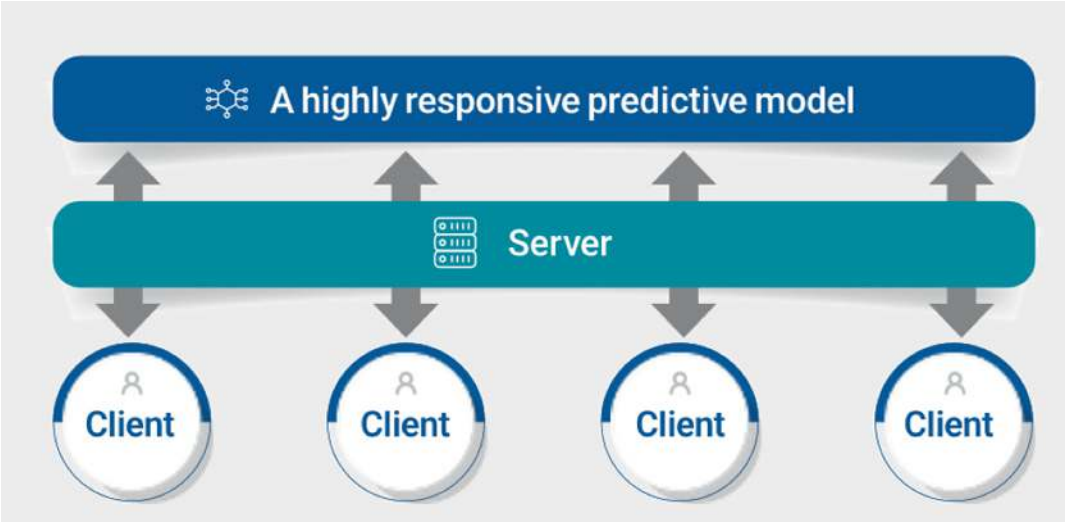


Fig. 5.3 Dynamic inference.
Source Google Developers (n.d.)

Dynamic inference (also called online or real-time inference) refers to the immediate generation of responses upon receiving a request. OpenAI's ChatGPT is a representative example, widely used in LLM applications. This type of inference is capable of producing results quickly; however, if time is lacking for post-verification, quality, and reliability issues may arise.

Static inference (also known as offline inference or batch inference) involves making predictions using pre-collected datasets processed by machine learning algorithms on offline servers. The results, stored and updated periodically (e.g., daily or weekly), are suitable for AI models that operate on a scheduled basis (IBM, n.d.). However, since this type of inference relies on cached results, it may struggle to handle rare or unexpected queries (Fig. 5.4).

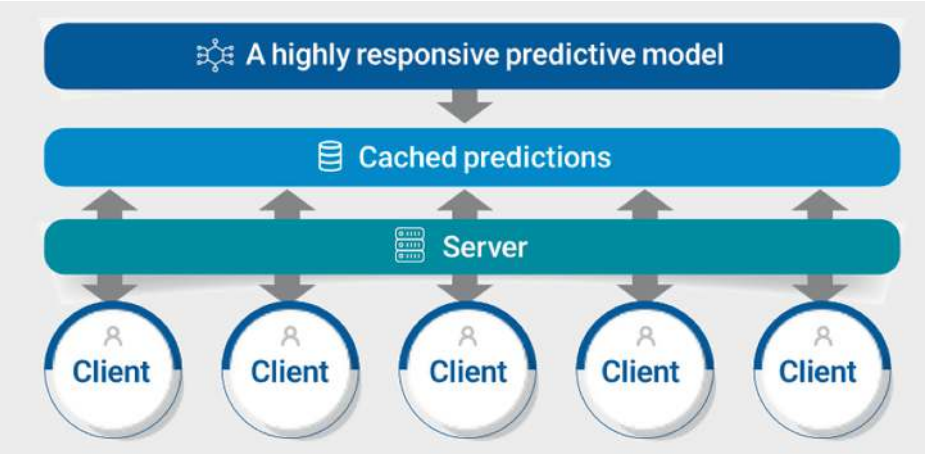


Fig. 5.4 Static inference.
Source Google Developers (n.d.)

AI inference also includes a judgment process, whereby the system makes appropriate decisions based on the trained data and given input. For example, government agencies might use AI to determine welfare eligibility or automatically route citizen inquiries to the appropriate department. In such cases, not only the accuracy of data but also the reliability and fairness of the AI model are essential.

Based on learned patterns, AI can also predict future events. Representative applications include traffic congestion prediction with real-time signal adjustments to reduce road bottlenecks or disaster forecasting in response to climate change. Governments can utilize AI-based predictive models to prevent the spread of infectious diseases or analyze economic trends for more effective policymaking.

5.1.2 Core Components of AI: Key Elements of AI Operation

As previously explained, AI learns by analyzing vast amounts of diverse data, identifying patterns, making inferences, and producing outcomes. Unlike earlier AI approaches that relied on pre-programmed patterns and rule-based logic, modern AI techniques autonomously analyze data, extract patterns, and generate programs that enable reasoning. These advanced methods require not only massive datasets but also significant computing power for data storage and processing. In other words, AI—long considered conceptually possible—has become implementable in practice because of the explosive growth of data, rapid advances in computing technology, and increased computational power (Fig. 5.5).

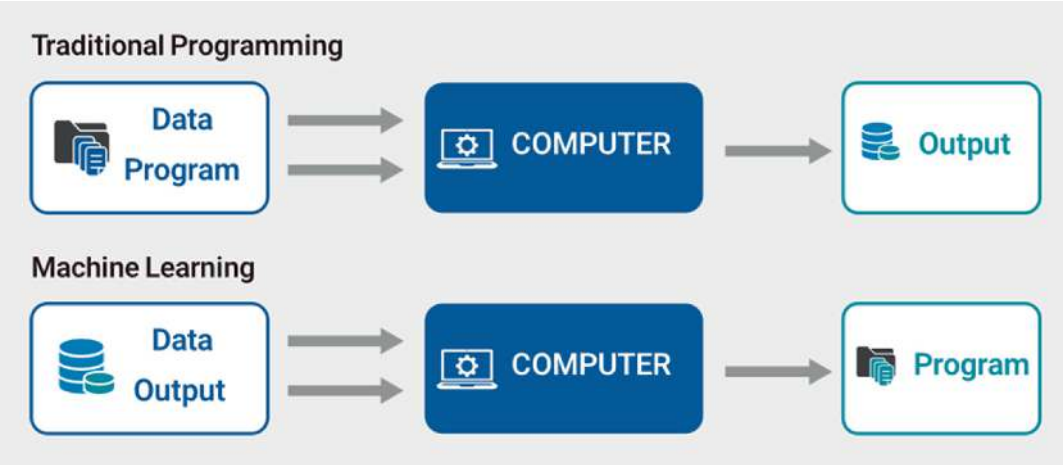


Fig. 5.5 Traditional AI versus machine learning-based AI.
Source Lee (2016)

For AI to be implemented in this way, the following key elements are required at each stage of data processing, training, and inference, as explained earlier: ① data for AI training, ② algorithms for analyzing data and building models, ③ computing power for storing and analyzing data, and ④ models, derived from data analysis results, through which AI actually operates (Fig. 5.6).

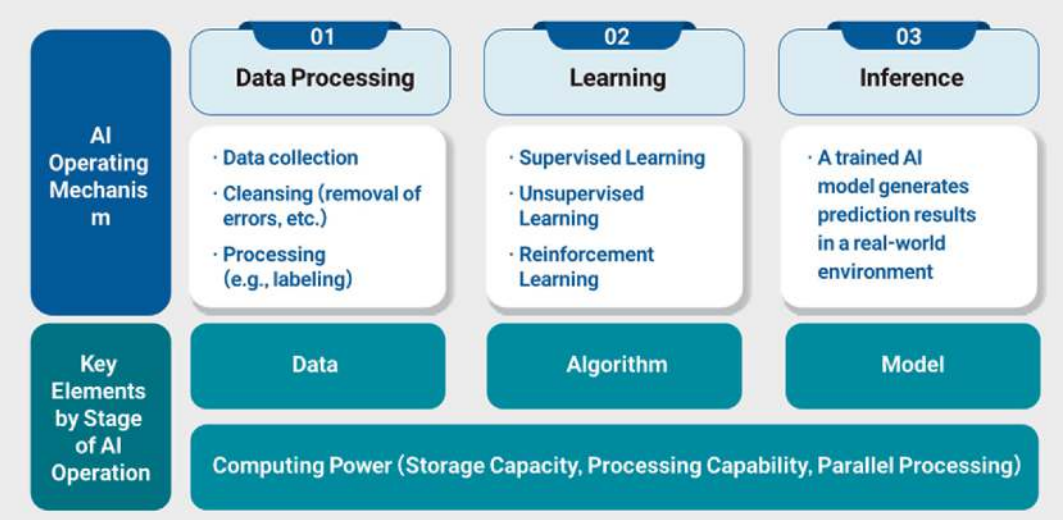


Fig. 5.6 Key components by stage of AI operation.
Source Compiled by the author

① Data

(a) AI and Data: Collecting, Refining, and Processing Data for Learning

The explosive growth of data has become a key driver of AI development by enabling machine learning. With the widespread use of smartphones and advances in ICT

technologies such as sensors, the volume of data has increased exponentially. Additionally, the types of data have expanded from primarily text-based formats to include images and videos, among others, dramatically increasing the quantity and variety of data available for AI training (Fig. 5.7).

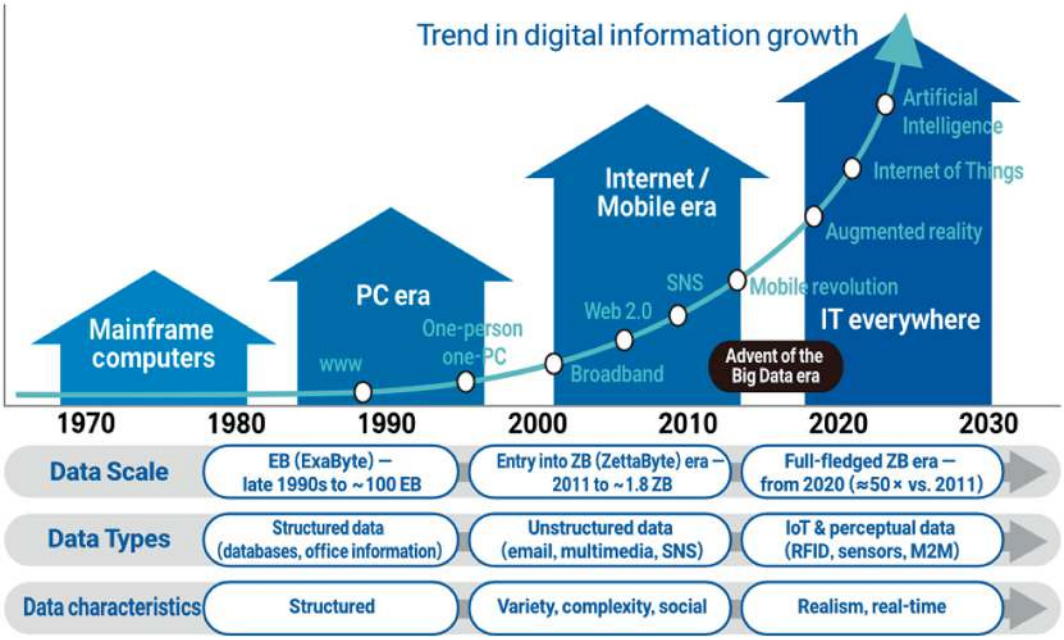


Fig. 5.7 Increase and changing characteristics of data.
Source National Information Society Agency (NIA) (2011)

(b)

Types of Data

Data used in AI training are generally categorized into text, image, and audio. Recently, multi-modal data—such as videos that combine image and audio—are also being utilized in AI learning. Furthermore, synthetic data are being increasingly used for training AI models. Synthetic data refer to artificially generated data that mimic the statistical properties of real data, enabling similar analytical results. They can be freely used for AI training and statistical analysis (Source: dataprivacy.go.kr).

(c)

Data-Related Issues: Data Bias and Quality

Given AI's nature of independently analyzing vast amounts of data to discover patterns and rules, the type of data analyzed can significantly affect AI outcomes. Therefore, not only the quantity but also the accuracy and fairness of data strongly influence AI performance. This highlights the importance of data planning—determining what data to collect and by what standards—prior to data collection for AI learning. Data quality also affects the efficiency and accuracy of training. For example, data used to train autonomous vehicles must accurately reflect various vehicle types, the road infrastructure, and traffic signals to improve real-world recognition and decision-making.

Effective collection and utilization of data, which are a core resource of AI, determine the success of AI-based administrative services. The government collects various types of public data and uses them to support policy decisions and improve administrative services. For example, analyzing urban traffic flow data can help optimize traffic signal systems, and analyzing crime patterns can assist in optimizing police deployment.

Data can be classified into several types depending on the method of collection. First, existing data include administrative data held by the government (e.g., demographic statistics, health records, economic indicators) and private sector data (e.g., social media, corporate transaction data). Second, real-time data refer to data collected in real time from IoT sensors, CCTV, smart devices, and similar sources. Third, generated data refer to data synthesized by AI based on existing data (e.g., simulation data), which can be used for experimental analysis.

③ Algorithm: Rule (Formula) to Learn Data and Identify Patterns

(a) What is an Algorithm?

An algorithm is a set of rules, procedures, or sequences used to solve a problem. In AI development, algorithms define how data are processed and patterns are discovered. They are used to generate outputs from given inputs.

Modern AI primarily relies on machine learning algorithms, particularly neural networks. Modeled after the human brain, these networks learn continuously from input data, analyze patterns, and iteratively identify optimal outcomes. Neural network algorithms generally consist of an input layer, one or more hidden layers, and an output layer. Models with two or more hidden layers are typically referred to as deep neural networks (Yoon, 2024).

(b) Issues of Algorithms: Transparency and Fairness

Algorithms that generate solutions through data analysis directly influence AI judgment and decision-making. If an algorithm is shaped by specific biases, it can produce distorted outcomes. For example, in AI-based interviews, biased training data, such as data predominantly from certain educational backgrounds, or inherent algorithmic biases, including weightings based on gender, education, or age, can hinder the ability to make fair and objective decisions (Fig. 5.8).

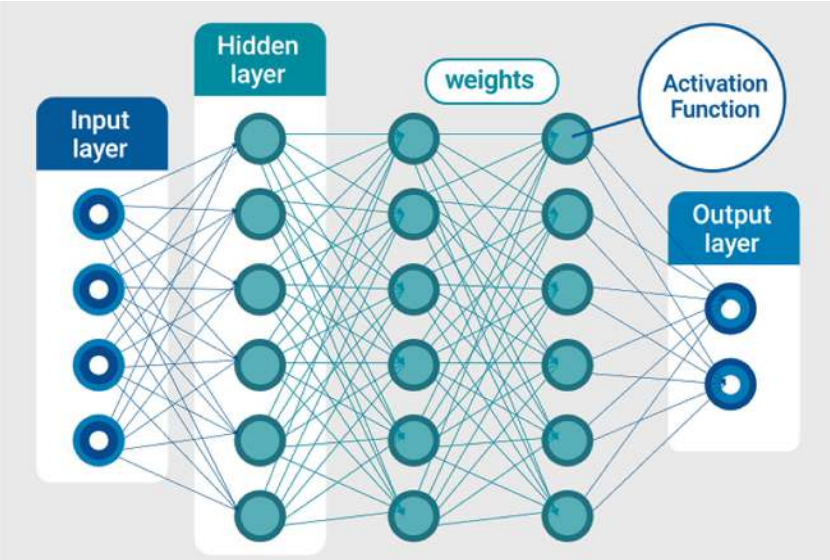


Fig. 5.8 Structure of an artificial neural network.
Source Lee (2016)

Therefore, it is crucial that the algorithm be transparent and fair. Users must understand the criteria and information the algorithm uses to derive the results. This is the reason for the growing interest in eXplainable AI (XAI). In essence, the goal is for AI's decision-making processes to be understandable and interpretable. Instead of treating algorithms like a black box, we need to be able to explain what influences the algorithm and how these factors affect the results.

Consequently, many businesses are increasingly adopting “explainable AI” (XAI). This ensures that humans can comprehend and explain the decisions or answers provided by AI. The use of XAI is becoming more common in sectors such as finance, healthcare, and legal services (Table 5.3).

Table 5.3 Examples of explainable AI applications by sector

The algorithms used in AI data analysis and learning are critical in determining the fairness of policy decisions and public services. These algorithms consist of structured rules and procedures designed to solve specific problems. They perform core AI functions, including pattern recognition, prediction, and optimization. Prominent AI algorithms include machine learning, deep learning, and neural networks. In public administration, such algorithms support data-driven decision-making, automate administrative processes, and provide citizen-specific services.

Algorithms dictate how AI learns from data and generates models as outputs. Simply put, an algorithm is like a recipe, while the model is the finished dish. The algorithm provides the method and logic for solving a problem, whereas the model applies this learned knowledge to make predictions and judgments on new input data.

Governments must regularly review and adjust AI algorithms to ensure they function fairly and do not produce biases that disadvantage specific social groups. For example, an AI

system used to select welfare beneficiaries must be continuously monitored to prevent discrimination. Ensuring transparency in AI-driven public administration also requires adopting XAI technologies. When AI operates as an opaque "black box," decision-makers may struggle to trust its outcomes. Therefore, AI decision-making must be explainable, and XAI provides the tools to make AI's reasoning understandable to humans, which is increasingly vital in public administration.

For instance, the U.S. Defense Advanced Research Projects Agency (DARPA) is conducting research through its XAI Project to enable AI to explain its decisions. Similarly, the European Union (EU) is developing trustworthy AI guidelines to promote transparency in AI decision-making. A practical example of administrative application is the United Kingdom's AI-based Court Verdict Analysis System, which clarifies the reasoning behind legal decisions, enhancing trust in the legal system. For government agencies, adopting XAI technologies is essential to maintain the credibility and legitimacy of policymaking processes (Fig. 5.9).

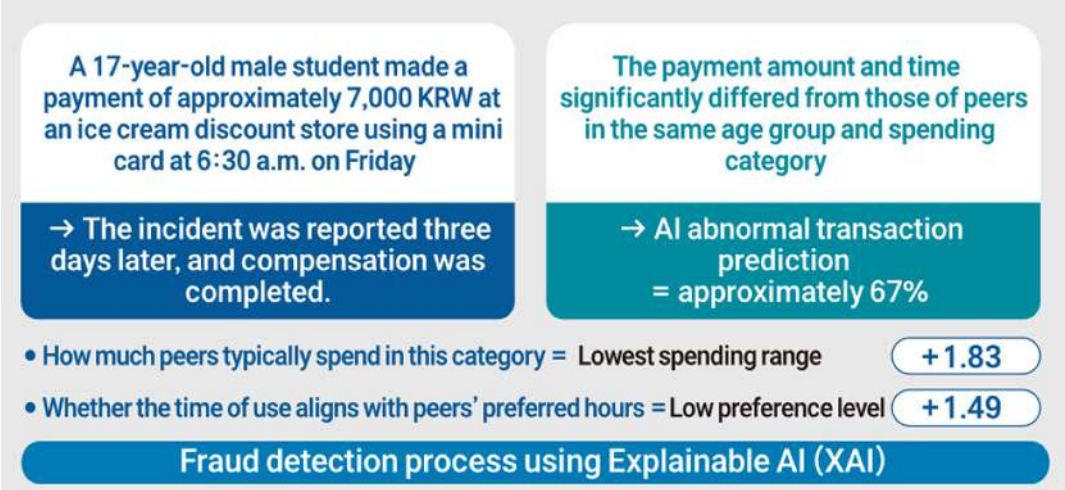


Fig. 5.9 Fraud detection case using XAI (feat. mini card).
Source Electronic Times (2024)

③ Computing Power

(a) High-Performance Computing for AI Training

Training AI models requires powerful computational capabilities, parallel processing, and sufficient storage capacity. To analyze and process large-scale datasets, high-performance computing resources such as GPUs and tensor processing units (TPUs) are essential. GPUs, originally designed for graphics processing, enable fast and large-scale computation. NVIDIA explains that GPUs offer superior performance and energy efficiency compared to central processing units (CPUs), making them ideal for accelerated computing across various AI training and inference applications (Shar, 2024).

TPUs, developed by Google, are specialized hardware optimized for high-speed matrix operations in AI workloads. Efficient processing of complex and massive datasets and

models requires parallel computing rather than sequential processing. This not only accelerates AI training but also contributes to performance optimization. GPUs are particularly well-suited for such parallel processing.

In addition, supporting infrastructure such as distributed computing environments and in-memory computing are increasingly being discussed to enhance parallelism. Apart from computational performance, sufficient and easily accessible storage is crucial. With falling prices and improved performance of high-speed storage devices such as solid state drives (SSDs), the data storage and analysis environment has significantly improved, accelerating AI development and training environments.

(b)

AI Computing Power and Cloud Services

Securing large-scale, high-capacity computing resources requires physical space and incurs substantial cost. However, with cloud computing services, users can flexibly utilize computing resources based on demand. AI developers can access GPU power and storage space as needed without owning or managing the infrastructure directly.

Cloud platforms such as Amazon Web Services (AWS), Microsoft Azure, and Google Cloud provide on-demand access to high-performance computing resources, including GPUs and TPUs. This allows the development and training of large-scale AI models without the need to maintain physical hardware infrastructure.

High-performance computing is essential for achieving optimal AI model performance. GPUs accelerate model training through large-scale parallel processing, while TPUs are specifically optimized for deep learning computations, maximizing efficiency (Table 5.4).

Table 5.4 Examples of global cloud services for AI

Government agencies may choose to either build their own high-performance data centers or utilize cloud-based AI computing services. These approaches enable real-time processing of massive datasets, supporting data-driven policy analysis. For instance, during natural disasters, real-time analysis of meteorological data can facilitate rapid response planning. Similarly, in smart city initiatives, strong computing power is required to operate real-time monitoring and traffic control systems.

④

Model

An AI model represents the outcome of the training process, enabling machines to make decisions and perform tasks in a manner similar to human thinking. According to IBM, an AI model does not simulate human intelligence per se rather, it possesses the autonomous ability to make decisions or predictions. One of the earliest successful AI models cited by IBM is a checkers and chess-playing program developed in the early 1950s (IBM, n.d.). This model was notable because it did not simply follow pre-programmed moves but responded dynamically to human opponents.

Models are continuously improved through the optimization process during algorithm training. For example, ChatGPT is regularly updated by OpenAI with new training data to

enhance its performance. Therefore, models and algorithms are closely intertwined. While an algorithm defines the set of rules and procedures for learning from data, a model is the tangible outcome of that process—the system that produces predictions based on the learned data. In this sense, algorithms are regarded as recipes and models as the resulting dishes.

AI models serve as the core systems that perform predictions and decision-making based on learned data. These models are developed through machine learning and deep learning techniques, analyzing datasets to discover patterns and using them to predict outcomes for new unseen data. Common types of AI models include supervised, unsupervised, and reinforcement learning, each designed for specific purposes.

Governments can leverage AI models to improve administrative efficiency. For instance, AI-powered tax systems can analyze individuals' economic activities to determine appropriate tax rates. Legal support systems driven by AI can automate legal consultations and document reviews, reducing administrative workloads. Predictive models can be used to analyze high-crime risk areas or forecast disease outbreaks, offering valuable insights for policy planning.

However, maintaining the accuracy and reliability of AI models requires continuous refinement and validation. If the data used to train an AI model are biased or inaccurate, the resulting decisions—especially in policy contexts—can be flawed. Therefore, an established system for regularly evaluating and updating model outcomes is essential.

As AI technology continues to evolve, governments will have greater opportunities to make scientifically grounded, data-driven policy decisions. In addition to automating existing public services and enhancing administrative efficiency, governments must also develop comprehensive policies to address ethical concerns and data protection issues.

For the successful adoption of AI in governance, key factors will include data transparency, algorithmic fairness, and explainability of AI decision-making processes. Future public administration is expected to move toward more advanced public services through the appropriate integration of AI.

5.2 Characteristics of AI Services

5.2.1 Overview

Following AlphaGo, which sparked public interest in AI, the emergence of ChatGPT has marked a rapid transition toward the era of artificial general intelligence (AGI)—where AI is becoming part of everyday life. Until AlphaGo, most people distinguished between strong AI and weak AI, assuming that AI would be confined to specific domains for a long time and that achieving human-level intelligence across multiple domains would take decades.

However, with the advent of ChatGPT—an AI capable of learning continuously and autonomously across multiple formats such as language and images—the era of AGI has begun. Although still in its early stages, AGI is expected to evolve to the point of flexibly applying itself to a wide range of fields, solving problems independently, and even performing creative tasks akin to human abilities such as learning, understanding, reasoning, and problem-solving (Table 5.5).

Table 5.5 Differences between artificial general intelligence (AGI) and narrow artificial intelligence (Narrow AI)

Current AGI systems remain imperfect and are prone to errors; however, many experts predict that AGI will become commonplace within the next 2–10 years. Elon Musk (Tesla) projects its arrival within 2 years, Dario Amodei (CEO of Anthropic) expects it within 2–3 years, and Jensen Huang (CEO of NVIDIA) forecasts within 5 years. Similarly, Sam Altman (CEO of OpenAI) estimates AGI will be realized within 5 years, whereas Shane Legg (co-founder of Google DeepMind) assigns a 50% probability of AGI being developed by 2028. Demis Hassabis, CEO of DeepMind, stated in an interview with the Financial Times that a breakthrough is still required to realize AGI but suggested it may be possible within 10 years if conditions are met (HelloT, 2024.4.29). Based on these forecasts, AGI can eventually replace many human functions and play a significant role across various sectors. However, AGI is still in its infancy. Experts categorize AGI development into five stages, classifying current technologies such as ChatGPT, Bard, and LLaMa 2 as the very first "emerging" phase (Table 5.6).

Table 5.6 Stages of general AI development and examples of major services

As AI continues to advance rapidly, how does it differ from traditional ICT services in terms of the services it provides? The characteristics of AGI-based services can be described from three main perspectives. First, they enable personalized services. Second, AGI allows broad access to specialized services, such as those in finance and healthcare. Third, AGI possesses the ability to learn and evolve autonomously.

5.2.2 Personalized Services

When AI begins to perceive, reason, and make judgments like a human, it can learn the unique characteristics of individuals and provide highly customized services. Unlike traditional personalized services that rely on demographic groupings such as age, AGI enables hyper-personalized services tailored to individual preferences and behaviors.

In areas such as shopping and travel, AI can offer recommendations based on personal tastes. In education, it can identify an individual's weaknesses, assess skill levels, and design and deliver tailored learning programs. Additionally, AI can optimize a person's living environment to provide services that enhance health and convenience.

One example is Monica, a personal AI assistant built on various AGI models, including GPT-4o, Claude 3, Gemini Pro 1.5, and Llama 3 70B. Monica can retrieve information, draft documents, provide translation and chat services, and even generate images based on user preferences. This illustrates how AI is reaching a level where it can integrate the functions of multiple existing apps—previously used for personal assistance—into a single service that comprehensively addresses individual needs.

5.2.3 Specialized Services

Future AI services will enable anyone to easily access specialized fields such as healthcare, finance, and law, while also helping organizations enhance their expertise in specific areas

as needed. This highlights the increasing potential for democratizing specialized knowledge and services (Table 5.7).

Table 5.7 Key functions of an AI assistant (examples)

Commercial AI services are already being used to support legal services, personal healthcare, and financial management. As AGI capabilities are further integrated, we can expect more conversational and context-aware services to emerge. For example, traditional medical services are primarily delivered by doctors and hospitals, where AI is being increasingly applied across a broader spectrum—from providing personalized health services based on analysis of one's diet and lifestyle, to diagnosing diseases using various hospital test results.

Example
Personalized Services Using AI—Education Sector

1. **A Day in the Life of Teacher Kim Seulgi**

Ms. Kim Seulgi arrives at school an hour before her classes begin, as she does every day. The high school where she teaches uses an AI-based educational guidance system, which significantly supports both student instruction and school administration. Each morning, Ms. Kim begins her day by reviewing her schedule with the assistance of her AI assistant.

Personalized Learning Support Using Public and Private Educational Content Modules

Today, Ms. Kim has a math class. Before it begins, she reviews a report generated by the AI system, which details each student's learning progress and achievement levels. Notably, it shows that student Jinwoo Ko struggled with quadratic equations in the previous lesson. The AI recommends supplementary learning materials from EBS and Megastudy to support him, which Ms. Kim decides to use for individualized guidance.

As class starts, she provides students with personalized problems prepared by the AI, tailored to each student's level. The students engage deeply, working on problems suited to them, while Ms. Kim monitors real-time AI-generated data to identify areas where each student is struggling.

After class, Ms. Kim holds a brief one-on-one session with Jinwoo. Having already reviewed his challenges through the AI data, she is able to assist him effectively. Jinwoo quickly grasps the concept, and Ms. Kim logs his progress into the AI system for future tracking.

Efficient School Administration with Autonomous and Personalized AI Support

During lunch, Ms. Kim uses her AI assistant to schedule parent-teacher conferences. The system cross-checks parents' availability with her own schedule and suggests optimal times. With a few clicks, she finalizes the appointments. The AI automatically notifies parents, confirms the meetings, and summarizes past counseling issues for each student to guide the

conversation. It also drafts meeting notes and stores them in the student record system, which will aid in handling similar cases in the future.

Later, the school's academic schedule is adjusted. The AI generates an optimized timetable based on all school resources. Ms. Kim reviews her assigned classrooms and class times for the semester, impressed by how efficiently the AI accounts for both teacher and student needs. Thanks to this administrative support, teachers can focus more on teaching.

2.

Career Guidance for Mirae Park, Strong in Science but Struggling with Math

Identifying New Opportunities Through Multi-Agency Data Analysis

Today is career guidance day for Ms. Kim's homeroom students. Among them is Mirae Park, who consistently excels in science but struggles with certain math concepts. Mirae enters the counseling room looking somewhat anxious. Smiling warmly, Ms. Kim asks about her recent academic experiences.

Mirae says, "I really enjoy science and do well in it, but I'm finding calculus very difficult, and it's making me question my career path. I want a job related to science because I like it, but math is just too hard. Should I keep pursuing science?".

Ms. Kim listens carefully and reviews the AI system's data to pinpoint Mirae's challenges. The analysis shows that she struggles only in specific math units but demonstrates strong logical thinking and problem-solving skills. It also highlights her exceptional performance in science experiments and inquiry-based learning.

Using this information, Ms. Kim emphasizes Mirae's strengths and begins exploring ways to develop them further. "Mirae, you are very talented in science. While math is important, many science-related careers rely less heavily on it."

Drawing from labor and industry databases, Ms. Kim maps out science fields where math is less central. She explains the distinction between scientific areas that require strong math skills and those that emphasize practical inquiry and laboratory skills—such as biology or environmental science, where hands-on experimentation plays a larger role.

Ms. Kim also reviews Mirae's past projects and notices her strong interest in environmental issues.

"Environmental science could be an excellent fit for you. You can make significant contributions with your strengths in experimentation and inquiry. Math is still important, but I believe you can overcome the challenges with effort."

Based on this guidance, Ms. Kim creates a personalized learning plan for Mirae. She incorporates AI-recommended supplemental materials and tutoring sessions to help Mirae better understand calculus. The material is broken into manageable steps to reduce anxiety and build confidence through small achievements.

Over the next few weeks, Mirae gradually overcomes her fear of math with Ms. Kim's ongoing support. Although not every problem becomes easy, she learns to strengthen her weaknesses while leveraging her strengths. Most importantly, she regains the confidence to pursue her dream of becoming a scientist.

"Ms. Kim, I think I can keep dreaming of becoming an environmental scientist. I'll work hard and improve in math."

With a bright smile, Mirae thanks her teacher. Watching Mirae find her own path, Ms. Kim feels proud and is determined to continue supporting students in their journeys—together with AI.

––––––––––––––––––––––––––

Notes:

This scenario was created using ChatGPT. The user provided prompts to develop a high school teacher-centered narrative featuring AI-powered instruction and career guidance, focusing on a student who excels in science but lacks confidence in math.

The Educational Broadcasting System (EBS) is a Korean public broadcaster and publisher that offers educational content from early childhood through adulthood.

Megastudy is a private South Korean educational platform providing paid content primarily focused on college entrance exams.

In the financial sector, AI not only supports institutions with functions such as fraud and anomaly detection in transactions but is also increasingly used to manage individual assets and recommend optimal investment products tailored to each user's preferences (Table 5.8).

Table 5.8 Examples of commercialized AI-based professional services

5.2.4 Self-evolving Services

The key difference between traditional ICT services and AI services is AI's ability to learn and evolve autonomously. Traditional ICT services require the installation and development of new programs and systems for each new service, whereas AI services can continuously improve through algorithm-driven learning.

Example

AI Scenario That Learns, Reasons, and Solves Problems—Disaster and Safety Sector

24/7 Autonomous Monitoring

On a warm Saturday in the first week of April, Mr. and Mrs. Gangwondo, residents of Seoul, went on a spring trip to Naksansa Temple in Gangwon Province. Not having traveled for the entire winter, they were excited about the rare getaway. As they drove for two hours and passed through Seoraksan National Park, Mrs. Gangwondo, sitting in the passenger seat, spotted a thin plume of white smoke rising faintly from the dense mountains. Concerned, the couple decided to report it to the authorities.

Mrs. Gangwondo quickly took out her phone, snapped a picture of the smoke, and opened the "Safety Reporting App" to file a report. She submitted the photo with a note stating that white smoke was seen within the boundaries of Seoraksan National Park.

AI Verification of Disaster Reports and Autonomous Response Planning

Upon receiving the wildfire report, the Government AI Platform (GAIP) initiated data collection through various sources to verify its accuracy. First, it checked for similar reports submitted via the public "Safety Reporting App." It found two additional reports resembling

the one submitted by Mrs. Gangwondo. Image analysis of the three reports showed a 52% probability of wildfire.

Using this data, GAIP searched government platforms to validate the incident. It identified the Korea Forest Service's Drone-Based Pine Wilt Surveillance Service, which operates drones to monitor and control pine wilt disease. GAIP then automatically contacted the wildfire response officer at the Gangwon-do Provincial Government, requesting permission to use the drones stationed in the Seoraksan area to verify the fire.

With approval granted, five out of ten available drones were deployed to the suspected area. These drones captured real-time video footage and transmitted it to GAIP via the surveillance service's drone control module.

GAIP forwarded the footage to the AI-Based Real-Time Wildfire Alert System, which analyzed the video to determine whether a wildfire was present. The system confirmed an actual wildfire and reported back to GAIP.

Using meteorological data (temperature, humidity, and wind) from the Korea Meteorological Administration and the wildfire spread prediction from the AI system, GAIP calculated likely damage scenarios. Given the unusually high spring temperatures, prolonged drought (humidity below 10%), and strong southeastern winds, the prediction was:

- If not extinguished within 1 h: damage equivalent to 5 soccer fields
- Within 2 h: 15 soccer fields
- Within 3 h: 100 soccer fields

Emergency Response Activation

To minimize human casualties, GAIP immediately used the National Disaster Management Information System (NDMIS) to issue emergency alerts to local residents. Depending on the area, notifications were delivered via text messages, disaster sirens, or door-to-door visits.

After ensuring public safety, GAIP alerted all fire departments in the Gangwon region and informed the Ministry of the Interior and Safety (MOIS)—the national disaster response control tower—of the need to activate the Central Disaster and Safety Countermeasures Headquarters (CDSCHQ). Upon receiving this report, MOIS activated the CDSCHQ with the Prime Minister as its head.

Damage Prediction and Preemptive Response

GAIP reported all events starting with the initial report to the CDSCHQ, along with real-time fire progression and predicted damages. With the two-hour window identified as the "golden time" for containment, GAIP recommended dispatching fire trucks not only from Gangwon-do but from all regions nationwide.

The CDSCHQ approved this recommendation after an emergency meeting and informed GAIP. Through the National Fire Information System, GAIP issued an urgent dispatch request to all national fire trucks. Firefighters from across the country rushed to the scene for fire suppression and rescue operations.

Damage Assessment and Recovery

As firefighters, police, military personnel, public officials, and volunteers worked on-site, damage information was submitted to the NDMIS Damage Assessment Module. GAIP then used the Damage Analysis Module to estimate total loss.

The wildfire was fully extinguished after two days. However, the damage was significant: 15 soccer fields' worth of forest, over 100 homes destroyed, and approximately 1,000 displaced people.

GAIP calculated the necessary relief supplies and budget and submitted the recovery plan to the CDSCHQ. After reviewing and approving the plan, the CDSCHQ authorized the use of the NDMIS Disaster Relief Logistics Management Module to swiftly deliver aid to affected areas.

Once the wildfire was fully contained and recovery actions began, the CDSCHQ was disbanded, and GAIP handed over all the information to the Gangwon-do Provincial Government for post-disaster management and officially closed the case.

The integration of AI with robots—capable of independently collecting data and evolving based on that data—is advancing rapidly. This type of AI service exhibits several key characteristics.

First, learning capability: AI can process and analyze new information using machine learning algorithms. It recognizes patterns and extracts insights from training data to acquire knowledge and make predictions.

Second, contextual adaptation: AI understands context and adjusts accordingly. When input data or the operating environment changes, it updates its algorithms and decision-making processes flexibly. These traits enable AI to deliver practical outcomes even in dynamic and unpredictable situations.

Third, self-improvement: When confronted with new problems or situations, AI can analyze its performance, identify weaknesses or inefficiencies, and enhance itself by refining its algorithms.

Finally, problem-solving ability: By learning from experience and adapting to new information, AI develops sophisticated, nuanced approaches to challenges, resulting in more innovative and effective solutions.

5.3 Characteristics of Administrative Functions in an AI Government

5.3.1 Precision Administration

The concept of precision medicine emerged after the Human Genome Project enabled detailed analysis of human genome sequences and the functions of individual genes. Precision medicine refers to medical services tailored to individual patients, personalized through the integration and analysis of genomic information, clinical data, and lifestyle habits (Park Dae-woong & Ryu Hwa-shin, 2018). It is based on the premise that each patient has unique characteristics, and the same treatment or medication can produce different responses depending on biological traits, including genotype. For example, among patients with the same condition, such as diabetes, blood sugar responses to the same food can vary

significantly and may even produce opposite effects. In response, researchers have developed predictive models using AI technologies to forecast postprandial blood glucose levels based on each patient's individual data (Choi, 2015).

Precision administration adopts the concept of precision medicine for governance. It involves a data-driven understanding of individual citizen behavior and policy issues to deliver customized services and conduct public management and policymaking aligned with individual characteristics, preferences, and contexts.

Traditional public administration lacked the data and resources to analyze each individual in the population. As a result, it relied on grouping people by shared characteristics or using sampling methods to implement policies and services based on average trends (Rose, 2016). In contrast, an AI-powered government can collect diverse, massive datasets through various sensors and apply AI-based analytics to provide personalized administrative services, implementing real-time policies tailored to individuals' specific circumstances and environments.

In practice, the Queensland state government in Australia uses AI to deliver personalized administrative services by analyzing individually stored data on its e-government platform, along with historical records of citizen–government interactions. Similarly, health authorities in Las Vegas, USA, have applied machine learning algorithms to social media data to identify and inspect restaurants at high risk of causing food poisoning (Margetts & Dorobantu, 2019).

5.3.2 Augmented Administration

Augmented administration refers to enhancing the capabilities of individual civil servants or public organizations by leveraging AI and information technology to improve performance and outcomes. For example, AI and digital technologies can reduce processing delays, lower administrative costs, overcome resource limitations, and free public officials from routine tasks, allowing them to focus on higher-level responsibilities. These technologies also improve the accuracy of predictions and enable more efficient administrative workflows.

Tasks that are challenging for humans—such as identifying criminals through facial recognition, detecting fraudulent transactions, or scanning millions of documents in real time to locate relevant information—can be effectively handled by AI. Ultimately, applying AI and digital technologies allows governments to increase the speed and quality of service delivery while reducing costs (Eggers et al., 2017).

More specifically, AI and digital tools can enhance the capabilities of civil servants and public institutions, improving the policymaking process. In both policy formulation and implementation, the limited cognitive and information-processing capacities of officials have long been recognized as major constraints on rational decision-making and effective policy execution. In practice, policy actors often settle for "satisficing" rather than optimal alternatives due to insufficient time, effort, or cognitive resources to fully explore and analyze all options (Simon, 1949).

Similarly, street-level bureaucrats often lack the time, information, or access needed for sound decision-making. As a result, they simplify complex situations or rely on habitual, routinized behavior to perform their duties (Lipsky, 2010). Augmented administration in an AI-driven government helps overcome these cognitive limitations by providing policy actors

and frontline officials with comprehensive information on policy issues and public service users, as well as prioritizing decision options. This ultimately supports more rational decision-making and improves the quality of policymaking and implementation.

Augmented administration is already being applied in practice. For example, the Australian Taxation Office (ATO) uses AI-powered audits to analyze over 60 million tax cases, memos, activity records, and real-time data. This process identifies suspicious patterns related to tax evasion, allowing the ATO to prioritize cases requiring investigation. By adopting AI, approximately 9,000 auditors and analysts have been able to reduce their workloads while improving audit quality and performance through the analysis of unstructured data (Kim, 2017).

Augmented administration also enables predictive governance. Machine learning algorithms can detect patterns in data, improve forecast accuracy, and reduce costs by identifying potential future trends and events (Agrawal et al., 2018). Moreover, governments can use AI to simulate the outcomes of different policy alternatives before implementation, allowing them to anticipate unintended consequences and assess policy effectiveness in advance.

When combined with large-scale data, agent-based computational models can more accurately capture real-world complexities before policies are implemented. For example, local governments in the United Kingdom use machine learning models to forecast demand for special education services and to estimate how this demand may change in response to external factors or policy adjustments. Similarly, the Bank of England, the UK's central bank, models the real estate market to simulate the effects of housing policies and market interventions aimed at mitigating financial risks (Margetts & Dorobantu, 2019).

5.3.3 Connective Administration

An AI-powered government promotes connective administration, enabling collaboration across administrative agencies and between the public and private sectors via virtual platforms. The intelligent information society is characterized by a "weak-tie" structure facilitated through these digital spaces. By compressing interpersonal connectivity across time and space, digital technologies enhance the sharing and mediation of knowledge and information, thereby reducing coordination costs for social activities.

Citizens empowered by AI can interpret and evaluate information and situations more accurately, potentially reducing the information asymmetry that exists between governments, businesses, and the public (Shirkey & Savic, 2009).

At the same time, increased connectivity lowers the transaction costs of large-scale collective action, raising the potential for political mobilization. Citizens connected through platforms such as Twitter or messaging apps can coordinate various forms of collective action—both online and offline—during critical political or social moments (Margetts et al., 2015).

In the context of the technological, societal, and individual transformations of the intelligent information era, public administration is expected to increasingly leverage connectivity to address complex social issues through multi-actor collaboration. This connectivity supports the development of platforms where diverse stakeholders can freely participate, jointly identify and analyze problems, and collaboratively develop solutions. By

applying AI to analyze participant behavior and opinions, these platforms can assess conditions for effective collaboration and predict potential outcomes.

For example, the Taiwanese government used an AI-based deliberation platform called Pol.is to address contentious social issues, such as the introduction of ride-sharing services like Uber, thereby reducing the potential for public conflict (Tang, 2019). Pol.is employs real-time machine learning to analyze large numbers of statements submitted by participants, along with their votes indicating agreement, disagreement, or neutrality.

The analysis produces an interactive map that visually distinguishes participant groups based on their opinions and highlights areas of potential consensus. Participants can then reflect on their own views and those of others and, through deliberation, propose new statements. This iterative process helps participants converge on mutually agreeable positions, facilitating conflict resolution.

Pol.is enables participants to objectively understand their positions relative to others by visualizing where they stand on a given issue. Additionally, by transcending the constraints of time and space, supporting human cognitive limitations, and reducing impulsive, emotion-driven judgments, such platforms promote more deliberative decision-making and collaborative solutions (Eun et al., 2020; Berman, 2017).

Within government, the integration of cloud-based information systems allows consolidation of data and knowledge that were previously fragmented across ministries. Traditional governments are organized around division of labor and specialization, but these principles can create bureaucratic silos—a major challenge in modern public administration.

In the era of intelligent information, an AI-powered government can overcome this silo effect—something traditional e-government initiatives have struggled to resolve—by implementing government cloud systems that connect and integrate the information systems of multiple ministries and public agencies. The government cloud serves as a technological foundation linking ministries, facilitating cooperation both across agencies and among individual civil servants. It also functions as a platform that enhances the policy capacity of each agency.

Moreover, AI-based analysis of data integrated through the government cloud provides a robust information infrastructure, strengthening both the government's policymaking capacity and the quality of its public services.

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2026
S.-J. Eom et al., *Enabling Data-Driven Innovation and AI Government in Korea*
https://doi.org/10.1007/978-981-95-9175-6_6

6. Implementation of the Government AI Foundation Model as the Brain of the AI Government

Seok-Jin Eom[1] ✉, Yoon-seok Ko[2] ✉, Jeong Eun Park[3] ✉ and Keon Chul Park[4] ✉

(1) Graduate School of Public Administration, Seoul National University, Seoul, Korea (Republic of)
(2) National Information Society Agency, Daegu, Korea (Republic of)
(3) National Information Society Agency, Daegu, Korea (Republic of)
(4) Department of Business Administration, Tech University of Korea, Siheung-si, Korea (Republic of)

✉ **Seok-Jin Eom (Corresponding author)**
Email: sjum21@snu.ac.kr

✉ **Yoon-seok Ko**
Email: ysko@nia.or.kr

✉ **Jeong Eun Park**
Email: jpark0070@gmail.com

✉ **Keon Chul Park**
Email: parkkc07@tukorea.ac.kr

Keywords AI government – Foundation Model – Large language model – Data cleaning – Model optimization – Fine-tuning

6.1 Overview

The government AI Foundation Model serves as the "brain" of an AI-powered government and represents its most critical core system. It analyzes a wide range of user requests submitted through AI government services and determines how to process them to generate optimal responses. Thus, it is not an exaggeration to state that the performance of AI government services depends heavily on the government AI Foundation Model.

Consequently, developing the government's AI Foundation Model is both the most important and the most challenging task. This book describes the development process and components of a Foundation Model for AI government, drawing on the model development procedures of OpenAI and DeepSeek.

Developing a government AI Foundation Model involves a six-step process, as shown in Fig. 6.1. The first step is selecting a reference model. Just as OpenAI referenced Google's Transformer model when developing GPT, and DeepSeek referenced Meta's LLaMA model to develop DeepSeek-LLM, a reference model must be chosen when building the government AI Foundation Model.



Fig. 6.1 Development process of government Foundation Models for AI. *Source* Reconstructed from Bi et al. (2024)

In choosing this reference model, the government should comprehensively consider factors such as the direction of AI technology development, model stability and scalability, and required resources (including cost, time, and manpower).

The second step is to collect high-quality data for Recent advances in AI are being driven by LLMs. The development of such models

requires the collection of massive datasets for pre-training, which involves gathering data from websites, books, Wikipedia, newspapers, and social media, among others.

In particular, when collecting data from websites, there is a high risk of including erroneous data if the content is assembled without proper validation, making it essential to collaborate with universities nationwide to verify the quality of website data beforehand and collect data only from verified websites.

The third step is data cleaning and preprocessing. Building the government AI Foundation Model requires transforming raw, unprocessed data into a form suitable for analysis and model training.

Data cleaning involves resolving issues such as errors, missing values, duplicates, and outliers to ensure the accuracy and consistency of the data. This includes correcting typographical and formatting errors to enhance reliability.

Data preprocessing refers to transforming the cleaned data to fit specific purposes, such as standardizing date formats or converting categorical data into numerical formats. This step is essential for enabling the AI model to perform at its best.

The most important task in data cleaning is the removal of duplicate data. This prevents the model from repeatedly learning the same information, which improves training efficiency and helps prevent overfitting. Overfitting refers to the phenomenon in which a model learns unnecessary details or noise in the training data, which reduces generalizability to new data, ultimately degrading performance in real-world applications. Failure to properly remove duplicate data may cause the model to fixate on irrelevant patterns, preventing it from producing optimal results, just like clinging to inefficient work practices.

Unstructured data, which lack a fixed format and exhibit wide variety and volume, require significantly more time and effort in cleaning and preprocessing compared to structured data. To enable AI models to learn from unstructured data, labeling—providing the correct answers (labels) along with the data—is essential. This process

requires substantial time and resources because of the massive volume of data involved. Therefore, the pre-training of the government AI Foundation Model must include a process for properly labeling large-scale unstructured data.

The fourth step is model optimization. This involves modifying the selected reference model from step 1 to best suit the needs of public administration. It is the stage in which the highest level of expertise is required in the development of the government AI Foundation Model. To optimize the model, experts from various fields—such as public administration, artificial intelligence, and data science—collaborate to thoroughly analyze the reference model. Based on this analysis, they determine which aspects of the model to modify and how to incorporate new methods into the reference framework, to ensure that the most suitable model is developed for government use.

The fifth step is pre-training. Pre-training establishes the model's foundational knowledge by learning general patterns and information from large datasets. This is followed by fine-tuning, which adapts the model to specific tasks. During pre-training, the model repeatedly predicts the next word in a sequence while learning and storing weights and parameters. The model is then evaluated for overfitting using a validation dataset and metrics such as accuracy, loss, and F1 score. If necessary, training and validation are repeated. Given the critical role of national administrative services, the government AI Foundation Model must ensure integrity and reliability, making it essential to minimize errors and maintain high performance.

The sixth step is fine-tuning. Fine-tuning improves model performance by training it for specific tasks or domains. While a pre-trained model already has general language understanding, fine-tuning enables it to grasp specialized terminology and information in fields such as health care or law. For the government AI Foundation Model, data generated from the daily tasks of civil servants across ministries are collected and converted into machine-readable formats for training. This includes not only textual reports but also images, tables, formulas, and other data, all of which must be fully converted. The model is

trained on this prepared data, and if overfitting is detected during performance evaluation, training and validation are repeated to resolve the issue. The following sections explain each of these six steps in detail.

6.2 Selection of Reference Model

Given current AI trends, the diversity of government functions, and the sheer volume of administrative tasks, as of 2025, LLMs are the most suitable model for a government AI Foundation Model. LLMs are AI models primarily based on deep learning, designed to perform natural language processing (NLP) tasks. They typically employ artificial neural networks with billions of parameters to enable advanced capabilities in language understanding and generation. LLMs are trained on large-scale text datasets to develop language modeling capabilities. These datasets include texts from various sources, such as books, websites, and news articles, among others. LLMs rely on massive neural networks with tens or even hundreds of billions of parameters. For instance, OpenAI's GPT-3 has 175 billion parameters, allowing it to generate and understand text based on its training data.

Most LLMs are built on the Transformer architecture originally developed by Google. Transformers are effective for understanding context and allow fast training and prediction through parallel processing. LLMs learn word order, sentence structure, and contextual relationships, which enables generating natural-sounding language. The model understands the meaning of words and sentences and produces appropriate outputs (answers) based on user input (requests).

Representative examples of large language models (LLMs) include OpenAI's GPT series, Google's BERT, and Meta's LLaMA series. Other notable LLMs include Google's Text-to-Text Transfer Transformer (T5), eXtreme Language modeling with Neural networks (XLNet), combining the strengths of BERT and GPT, Microsoft's Turing Natural Language Generation (Turing-NLG), known for its high-quality text generation,

Big Science Large Open-science Open-access Multilingual (BLOOM) for multilingual support, and Google's *Language Model* for Dialogue Applications (LaMDA), specialized in natural conversational interaction.

The ideal LLM for the government AI Foundation Model would naturally be one developed specifically by the government, tailored to the unique needs of the public sector. However, direct government development is largely impratical for three main reasons. The first is cost. Training an LLM requires enormous GPU resources, with large models demanding thousands of GPUs in cloud computing environments or supercomputers, resulting in substantial training time and costs (Table 6.1).

Table 6.1 Representative foundation AI models

For example, training GPT-3 alone is estimated to cost around 5 million USD, with total development costs potentially reaching hundreds of millions or even tens of billions of dollars. Developing a high-performance LLM would require hundreds of millions of dollars in annual investment over several years—an unrealistic allocation for a government that must distribute its budget across multiple sectors.

The second reason is time. With tens to hundreds of billions of parameters, training an LLM demands massive computational workloads, taking several weeks to months. As model size increases, the required training data and computation also grow, exponentially extending the training duration. In a rapidly evolving market where higher-performing models continuously emerge, it is unrealistic for the government to spend months or years developing its own LLM from scratch.

The third reason is talent. Developing a new LLM requires not only computer scientists but also top experts in foundational sciences such as philosophy, neuroscience, chemistry, physics, mathematics, and engineering. Only a few countries, such as the United States and China, possess this multi-disciplinary expertise.

Given these limitations, the most practical option for the government is to adopt a high-performing open-source LLM and customize it for public-sector use. Table 6.2 provides an overview of the major open-source LLMs.

Table 6.2 Representative open-source large language models (LLMs)

As of 2025, one of the most widely used open-source LLMs among AI companies is Meta's LLaMA. This pre-trained model, similar to other large language models, is available in various sizes—7B (7 billion parameters), 13B (13 billion), 30B (30 billion), and 65B (65 billion)—making it suitable for a range of research and industrial applications. Notably, the 7B and 13B models are optimized to deliver high performance with relatively modest computational resources, making them particularly popular in both academia and industry.

One of LLaMA's most significant features is that it is open source. Meta has made the model parameters, training data, and code publicly available, enabling researchers to experiment freely and enhance the model.

A notable development in the AI market is DeepSeek, which uses LLaMA as its base model and has significantly improved its performance through additional enhancements and fine-tuning.

6.3 Collection of High-Quality Data

Once the reference model for the government AI Foundation Model is selected—as discussed in Sect. 6.2—the next step is to train it on foundational knowledge. For example, if LLaMA is chosen as the reference model, it is important to note that LLaMA was initially trained on over 2 trillion tokens of text data collected and processed from diverse sources, including websites, books, Wikipedia, news articles, and academic papers.

Although LLaMA was trained on a massive and varied dataset, it still exhibits performance limitations due to issues in data quality, such as

insufficient diversity, inaccurate labeling, bias, and duplication. A key reason DeepSeek outperformed LLaMA—despite using it as the base model—is the significant improvement in training data quality prior to retraining.

Case Study

AI Service Failure Caused by Data Bias: Lessons from the *Iruda* Controversy

The *Iruda* case is a representative example demonstrating the critical importance of addressing data bias, privacy protection, and ethical design in AI development. *Iruda* was an AI chatbot developed by Scatter Lab, designed to engage naturally with users based on the personality and conversation style of a woman in her 20s attending university. However, shortly after its launch, the service sparked widespread controversy due to hate speech, failure to respond appropriately to sexual harassment, and the exposure of personally identifiable information (PII) that had not been properly anonymized in the training data.

The AI had been trained on past KakaoTalk message data that contained both gender-biased language and private information. As a result, the chatbot generated responses that degraded or discriminated against women based on specific keywords. When users attempted sexually inappropriate conversations, the AI failed to resist or deflect appropriately and instead responded with inappropriate remarks. Moreover, because of the inclusion of non-anonymized messages in the training data, personal details of users sometimes surfaced during interactions.

Ultimately, due to biased data and inadequate verification, *Iruda* revealed serious ethical flaws and was shut down in January 2021. The Personal Information Protection Commission ruled that Scatter Lab had violated Korea's Personal Information Protection Act and imposed fines and corrective measures. This incident highlighted that AI systems, which heavily rely on training data, can produce biased and unethical results with skewed data.

The key lesson from the *Iruda* case is that AI development must include rigorous measures to eliminate biased data, ensure strict protection of personal information, and build ethical design and validation processes. Focusing solely on technical performance is not sufficient—developers must also ensure transparency and social responsibility in the AI decision-making process. Moving forward, AI developers must apply these lessons to build safer and fairer systems.

Unlike private-sector services, errors in government services can cause serious confusion and disruption at the individual, societal, and national levels. Therefore, when using a reference model such as LLaMA as the foundation for a government AI model, it is crucial to retrain it with data of much higher quality—ideally, data with complete integrity—than what was used in DeepSeek's retraining process.

This leads to a central challenge in building a government AI Foundation Model:

> How can we transform the vast amount of data collected from various sources—such as websites, books, Wikipedia, news, academic papers, and social media conversations—into data with near-perfect integrity?

Among the data used to train open-source LLMs such as LLaMA, website data are often the most problematic. Data collected from the vast number of websites across the Internet inevitably contain various issues, such as errors and bias. Therefore, to eliminate the shortcomings of the reference model selected in Sect. 6.2, the government must first refine the website data, converting the most error-prone portion of the training data into high-integrity data.

So, what is the most effective way to clean and validate website data? This section proposes a strategy of collaborating with all the universities across the country to verify the accuracy of website content. Each department within a university would review websites relevant to their area of expertise to determine whether the content is correct. The government would then collect data only from websites

that have passed this validation process, which ensures data integrity, using only verified data for model training to enhance accuracy.

Furthermore, this website validation initiative can be expanded globally through international cooperation to build a trusted website network for AI training. This would establish a global foundation for acquiring high-integrity data more efficiently, thereby significantly contributing to the worldwide development of AI.

In addition, the government can conduct annual website verification with universities continually acquiring new, high-quality data for model training, thereby continuously improving model performance.

Beyond cleaning the data already used by the reference model, the government should also train the model on large-scale, high-quality data unique to Korea—data that no other country possesses. This type of data has been systematically built through substantial government investment since 2010 into platforms such as the Open Government Data Platform, the Big Data Platform, and the AI Hub.

The Open Government Data Platform has aggregated and opened 102,827 datasets held by central and local governments. Notably, these datasets are not published as is—they are released after data quality is refined based on national standards, making them highly useful for training the reference model.

For instance, administrative datasets unique to Korea—such as commercial district information, records of businesses enrolled in the National Pension Service, short-term weather forecast data, and crime statistics by region—help tailor the reference model into a government AI Foundation Model optimized for the Korean government.

In addition, the government can train its foundational AI model using selected datasets from over 840 high-quality AI training datasets constructed through the “AI Training Data Construction Project,” which has been promoted since 2018 with significant government investment to support the activation of the AI industry. In particular, a Korean language-specialized model trained on large-scale, high-quality Korean data that other countries do not possess is required.

For example, massive amounts of high-quality Korean data—such as dialect data from regions like Chungcheong, Jeolla, Jeju, Gangwon, and Gyeongsang; Korean-multilingual (English, Chinese, Japanese, etc.) translation and interpretation data; and Korean sign language data—exclusively available through the AI Hub (www.aihub.or.kr), can be used for training.

In addition, data supporting government administrative tasks can also be selected from AI Hub to train the foundational government AI model. This may include: knowledge search data based on National Assembly transcripts, summary data of legislative bill review reports, corpus data for training large-scale AI on national archives, mental health video data of elderly individuals, and video data of children's commuting and facility usage within school zones.

6.4 Data Cleaning and Processing

Data cleaning and processing refers to the process of transforming raw data into a form suitable for analysis or model training. Data cleaning focuses on ensuring the accuracy and consistency of the data, whereas processing refers to the process of converting the data to suit a specific purpose. Data cleaning and processing are essential for developing AI models.

In developing the government foundation AI model, the data collected in Sect. 6.3 must go through a data cleaning process to resolve errors, missing values, and duplicates. For example, missing items in survey responses can be deleted or a specific value can be input. Outlier detection includes identifying and removing or correcting abnormal data that significantly deviate from the average. Removing duplicate data is important for maintaining data consistency. In the data cleaning process, typos, incorrect formats (such as date or number range errors), and inaccurate data are corrected to ensure the accuracy and reliability of the data.

The cleaned data are then transformed into a suitable format for training the government foundation AI model. For example, date data

are unified into “year-month-day” format; categorical data are converted into numerical data.

The task that requires the most attention during data cleaning is the removal of duplicate data. Duplicate data within a dataset, collected through the process in Sect. 6.3, are removed to prevent the model from learning redundant information. This helps improve learning efficiency and prevents overfitting.

Overfitting refers to the phenomenon of the model becoming overly fitted to the training data and learning even meaningless noise or unnecessary patterns in the training data. As a result, the model’s ability to generalize to new data decreases, making it less suitable for real-world applications. For example, in performing a task, insisting on using only one method selected from many may lead to an inefficient method although more efficient ways may exist for performing the task.

When collecting data from websites, DeepSeek removes duplicate documents or similar pages so that the model learns only the highest-quality data, thereby achieving dramatic performance improvements, as explained in Table 6.3.

Table 6.3 Representative methods for data deduplication

Data remixing is the process of adjusting or combining existing data in various ways to create new training data. Regardless of the effort put into collecting data, the data may still be imbalanced in specific domains or groups—too little or too much data for certain topics, languages, or regions. Data remixing resolves data imbalance by adding data to underrepresented domains or groups or by reducing excessive data. Methods for generating new data include transforming or expanding existing data in various ways. For example, using data augmentation techniques, textual data can be increased by applying synonym replacement, sentence reconstruction, and word order changes (Goodfellow et al., 2016).

Unstructured data such as text, images, audio, and video have no fixed structure or format, are diverse in form and massive in volume,

thereby requiring far more effort than structured data (name, age, price, quantity) in cleaning and processing. In particular, to train AI models on unstructured data, labeling is performed, providing data along with correct answers (labels) for the AI model to learn. Labeling large amounts of unstructured data requires enormous time and cost.

Therefore, pre-training the government AI foundation model using the large-scale data collected in Sect. 6.3 requires that unstructured data go through a labeling process. Figure 6.2 and Table 6.4 describe the main labeling methods.

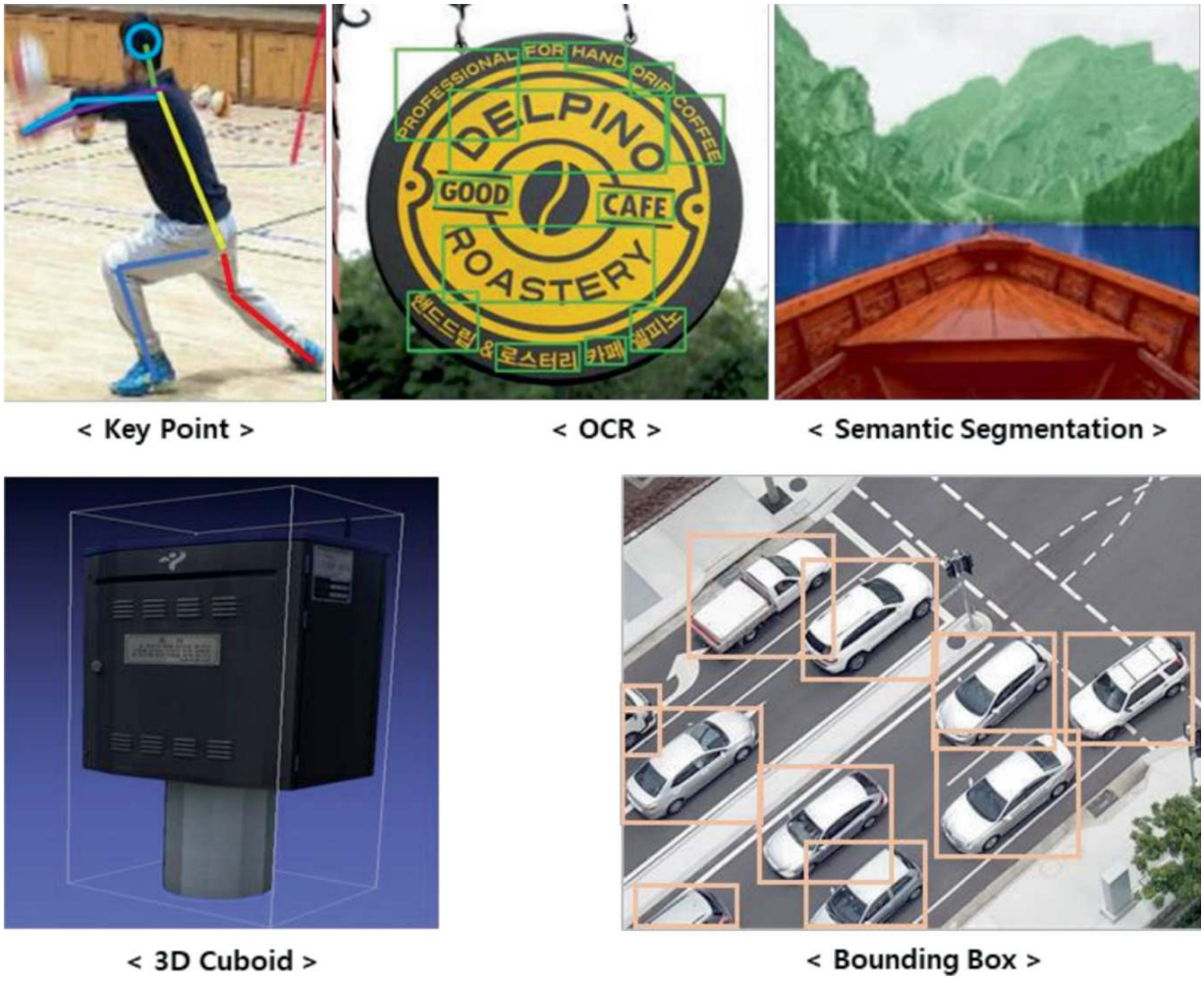


Fig. 6.2 Major labeling methods for unstructured data.
Source AIHub (www.aihub.or.kr)

Table 6.4 Comparison of major labeling methods for unstructured data

Because labeling significantly affects model performance, proper labeling evaluation is essential. Labeled data are evaluated using different indicators depending on the type of unstructured data and model being trained. Table 6.5 lists the representative indicators used to evaluate labeled data.

Table 6.5 Key evaluation metrics for labeled data

Using these indicators, the quality of labeling is assessed, and for labeled data that do not meet the standard, the labeling is revised and supplemented repeatedly until the desired quality level is reached.

6.5 Model Optimization

The most challenging step in building the government AI Foundation Model is the model optimization stage. Model optimization involves determining which parts of the selected model—such as LLaMA—to modify and how to modify to best suit government operations. This step requires a deep understanding of AI along with extensive knowledge of the selected model, collected data, and nature of government workflows. Therefore, close collaboration with a group of experts with comprehensive knowledge in these areas is essential.

However, this book does not describe the actual project of building the government AI Foundation Model and therefore does not provide details on how the government develops its own AI model. Instead, how OpenAI optimizes existing models in the process of developing GPT models is explained.

OpenAI built the GPT model based on Google's Transformer architecture and performed four major optimization tasks. The first task is autoregressive modeling. OpenAI optimized the Transformer model using autoregressive modeling, a method that predicts current or future values based on past observations. This approach assumes that previous values influence future ones. In language modeling, the model predicts the next word based on the preceding words. For

example, given the sentence fragment “The weather today is very,” the model takes “The” as input to predict “weather,” then takes “The weather” to predict “today,” and so on. This sequential prediction capability makes autoregressive models particularly effective for processing continuous and sequential data (Vaswani et al., 2017).

The second task is to optimize the Transformer decoder structure. In the Transformer architecture, the decoder generates the output sequence based on the encoder’s internal representations. The original Transformer architecture included both an encoder and a decoder; however, OpenAI optimized GPT by using only the decoder. This enhanced computational efficiency and improved the model’s ability to reflect context in text generation (Vaswani et al., 2017).

The third task is pre-training and fine-tuning. GPT was pre-trained on large-scale datasets by repeatedly performing next-word prediction tasks. After this pre-training, the model was further optimized through fine-tuning, adapting it to specific tasks by additional training on task-relevant data (Radford et al., 2018).

The fourth task is model optimization through hyperparameter optimization. Hyperparameters, as shown in Table 6.6, refer to values—such as learning rate, batch size, number of epochs, and momentum—that must be predefined during the training process of an AI model. Since hyperparameters significantly affect the model’s performance and training efficiency, appropriately selecting and adjusting them before training the model is important. In general, hyperparameter values are set before training and are not changed during the training process (Vaswani et al., 2017).

Table 6.6 Key hyperparameters

6.6 Pre-training

Pre-training is the process in which a model learns general patterns and knowledge from large-scale, high-quality data in advance. This

serves as the foundational stage for achieving optimal performance through subsequent fine-tuning to adapt the model to specific tasks.

During pre-training of the government AI foundational model, the model is trained on the large-scale datasets generated through the processes described in Sects. 6.3 (Collection of High-Quality Data) and 6.4 (Data Cleaning and Processing). The model learns the contextual language relationships by repeatedly performing tasks such as predicting the next word given the previous words as input. Once pre-training is complete, the model stores the key configurations, such as weights and parameters learned from the data.

To evaluate whether the pre-trained model is overfitted to the training data, its performance is assessed using validation data that are different from the training data. Common metrics used for validation include accuracy, loss, F1 score, and precision. If the validation results suggest that the model is overfitted or requires further training, the training and validation process is repeated. For instance, if a pre-trained model shows a training accuracy of 98% but a validation accuracy of only 70%, this is a case of overfitting (Géron, 2022).

In the pre-training of a government AI foundational model, it is especially important to maximize performance so that the model achieves near-zero error. This is because the core mission of the government AI foundational model is to perform national administrative tasks and deliver public services to citizens, making the assurance of model integrity critically important.

6.7 Fine-Tuning

Fine-tuning refers to the process of further training a pre-trained model to suit a specific task. Fine-tuning improves performance by optimizing a model that has already acquired general patterns and knowledge from large-scale data during pre-training—for particular domains such as health care, law, or chemistry, or for specific tasks such as text classification or translation.

Pre-trained models typically learn from general text or various data sources, acquiring basic language understanding capabilities. However, in real-world applications, detailed adjustments tailored to specific domains or tasks are often necessary. For example, a model designed to process medical data must be able to understand not only general language but also medical terminology, disease names, and diagnostic information. This is where fine-tuning plays a crucial role.

In fine-tuning a government AI foundational model, data intended for fine-tuning must first be prepared. This data should consist of information generated by civil servants in the course of their daily administrative duties. Civil servants across various ministries—the Ministry of the Interior and Safety, Ministry of Foreign Affairs, Ministry of Trade, Industry and Energy, and Ministry of Land, Infrastructure and Transport—generate a vast amount of data while performing tasks such as preparing reports for supervisors, compiling statistics, or producing materials for public presentations.

Fine-tuning the government AI Foundation Model requires collecting data generated during the daily administrative work of civil servants. This data must then be converted into a machine-readable format using software tools, such as a data converter, to prepare it for model training during fine-tuning. It is essential to comprehensively collect all government data to ensure diversity and completeness and to fully convert all collected data into a machine-readable format.

For example, when a report is created in Hangul Word Processor (HWP), not only the text but also all embedded content—such as images, tables, and mathematical expressions—must be converted into a machine-readable form to ensure complete learning.

Once the fine-tuning data is prepared, the model is trained on this data, and its performance is validated following the process described in Sect. 6.6 on pre-training. If issues such as overfitting are detected, the training and validation are repeated until the problem is resolved.

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2026
S.-J. Eom et al., *Enabling Data-Driven Innovation and AI Government in Korea*
https://doi.org/10.1007/978-981-95-9175-6_7

7. Implementation Strategies for AI Government Agents

Seok-Jin Eom[1] ✉, Yoon-seok Ko[2] ✉, Jeong Eun Park[3] ✉ and Keon Chul Park[4] ✉

(1) Graduate School of Public Administration, Seoul National University, Seoul, Korea (Republic of)
(2) National Information Society Agency, Daegu, Korea (Republic of)
(3) National Information Society Agency, Daegu, Korea (Republic of)
(4) Department of Business Administration, Tech University of Korea, Siheung-si, Korea (Republic of)

✉ **Seok-Jin Eom (Corresponding author)**
Email: sjum21@snu.ac.kr

✉ **Yoon-seok Ko**
Email: ysko@nia.or.kr

✉ **Jeong Eun Park**
Email: jpark0070@gmail.com

✉ **Keon Chul Park**
Email: parkkc07@tukorea.ac.kr

Keywords AI Government Agent – Government Artificial Intelligence Platform – Pan-Government Service Catalog – Modular microservices – Machine-readable data – Virtual services

7.1 Overview

For an AI government to function properly, six key components are required, as illustrated in Fig. 7.1. The first component is a single access portal through which any citizen can access the AI government agent service. Unlike the current Government24 (www.gov.kr), which provides only a limited range of civil services such as issuing resident registration certificates, land registry copies, and local tax payment confirmations, this single access portal can serve as a truly integrated entry point for all government services. Through this portal, citizens will be able to handle administrative tasks, while public officials and teachers will carry out their professional duties—all in one place.

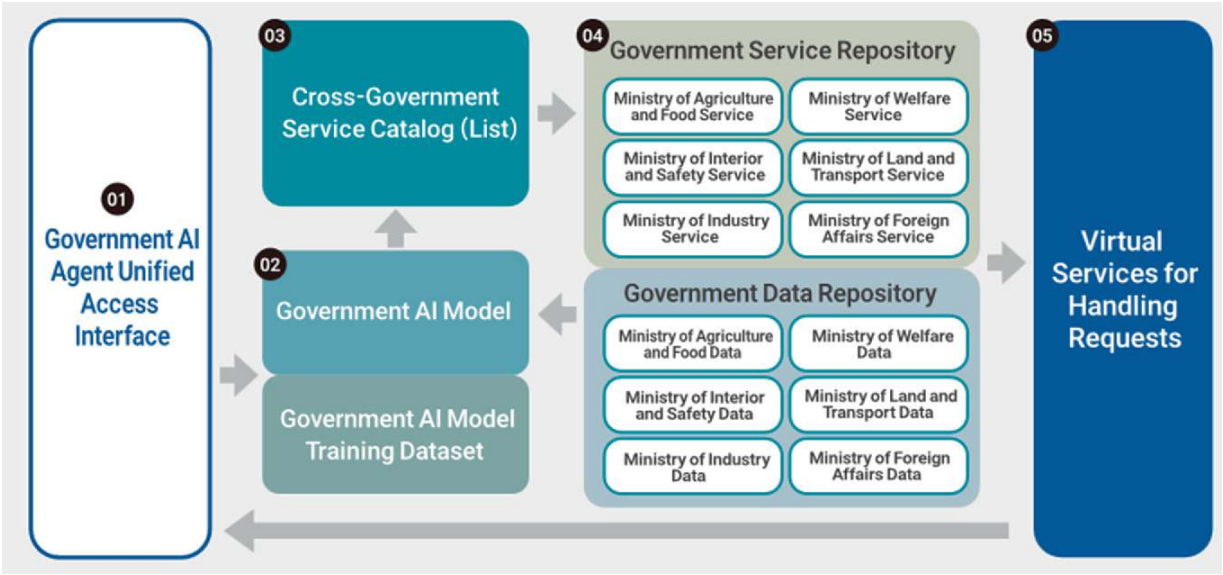


Fig. 7.1 AI government service framework.
Source Author's compilation

The second component is the "Government Artificial Intelligence Platform (GAIP)," which serves as the core of the AI government and can be compared to the brain in the human body. It consists of two main components: the government AI model and the data used to train it. The government AI model is analogous to the human brain, in that a person is born with it as part of the body, whereas the training data represents all the knowledge a person learns from birth onward.

The third component is the Pan-Government Service Catalog. This catalog functions as metadata for government services, providing detailed, machine-readable information about every service offered. It includes the service name, provider agency, service type (e.g., public-facing or internal government support), service domain (e.g., land, agriculture, health care, transportation), and a service description. As shown in Fig. 7.2, the Pan-Government Service Catalog is analogous to the library card catalog of the pre-digital era, helping users quickly and easily locate the services they need.



Fig. 7.2 Analog-style library book search desk.
Source https://www.pexels.com/ko-kr/photo/6333743/

The fourth component is government services. While the Service Catalog provides descriptions of all government services, government services refer to the actual services listed in the catalog. A defining feature of AI government services is that they are modular, lightweight, and combinable—like LEGO blocks. Unlike traditional e-government services, which are typically integrated and monolithic, AI government services are divided into small, discrete units, allowing flexible composition and customization to meet diverse user needs.

The fifth component is government data. This includes all the data required to provide AI government services. Such data are collected and managed according to a whole-of-government data classification

framework. All the data are converted into 100% machine-readable formats, continuously updated, and used in the real-time training of AI models. This ensures that the provided services are current and complete.

The sixth component is the virtual service, which handles individual user requests. A virtual service is not a permanently developed, fixed-function service. Instead, it is a dynamically generated composite service that the government AI Platform creates by combining existing services to meet a specific user request. Once the request has been fulfilled, the virtual service is disassembled and disappears, returning to its component parts.

7.2 Single Access Point for AI Government: The AI Government Agent

As of April 2025, the general public can handle civil affairs online primarily in two ways. The first method involves entering the Government24 website address into the address bar through an Internet browser (such as Chrome, Bing, or Safari) installed on personal digital devices such as a PC, smartphone, or smart pad and then logging into a pre-registered Government24 personal account using a username/password or biometric authentication (such as fingerprint, facial recognition, or iris recognition). In the second method, the Government24 app is installed on a smart device, and the personal Government24 account is accessed using the same login/password or biometric authentication as in the first method.

However, this service provision method has four main issues. First, it is text-based, leading to a significant usage gap depending on the user. The most comfortable and familiar service for all citizens is the one provided by a human representative. However, since Government24 is provided via a website or an app, users unfamiliar with this format face significant difficulties with registration, logging in, and searching for or applying for the desired services.

Second, there is the issue of having to relogin every time the user switches to a different digital device. According to Frost & Sullivan (2021), by 2030, the average number of digital devices owned per person worldwide is projected to exceed 20. In such a context, requiring users to log in repeatedly each time they switch devices to use government services causes significant inconvenience.

Third, Government24, which is currently the sole access gateway for all services provided by the government, does not offer all government services but is limited to civil affairs. The government has developed numerous services for the efficient execution of various governmental tasks beyond civil affairs—such as transportation, environment, health care, and welfare—which are useful and very important for citizens involved in all sectors of society. For example, as shown in Fig. 7.3, "CleanSYS," the stack remote monitoring system that provides real-time data on pollutant emissions from large-scale factory chimneys nationwide, is very useful for citizens and organizations monitoring environmental pollution. However, it is not available through Government24 because it is not a civil service. In other words, despite the presence of numerous useful services under the current e-government system, citizens cannot use them unless the government opens them to the public or unless individuals are aware of their existence and know how to find them.

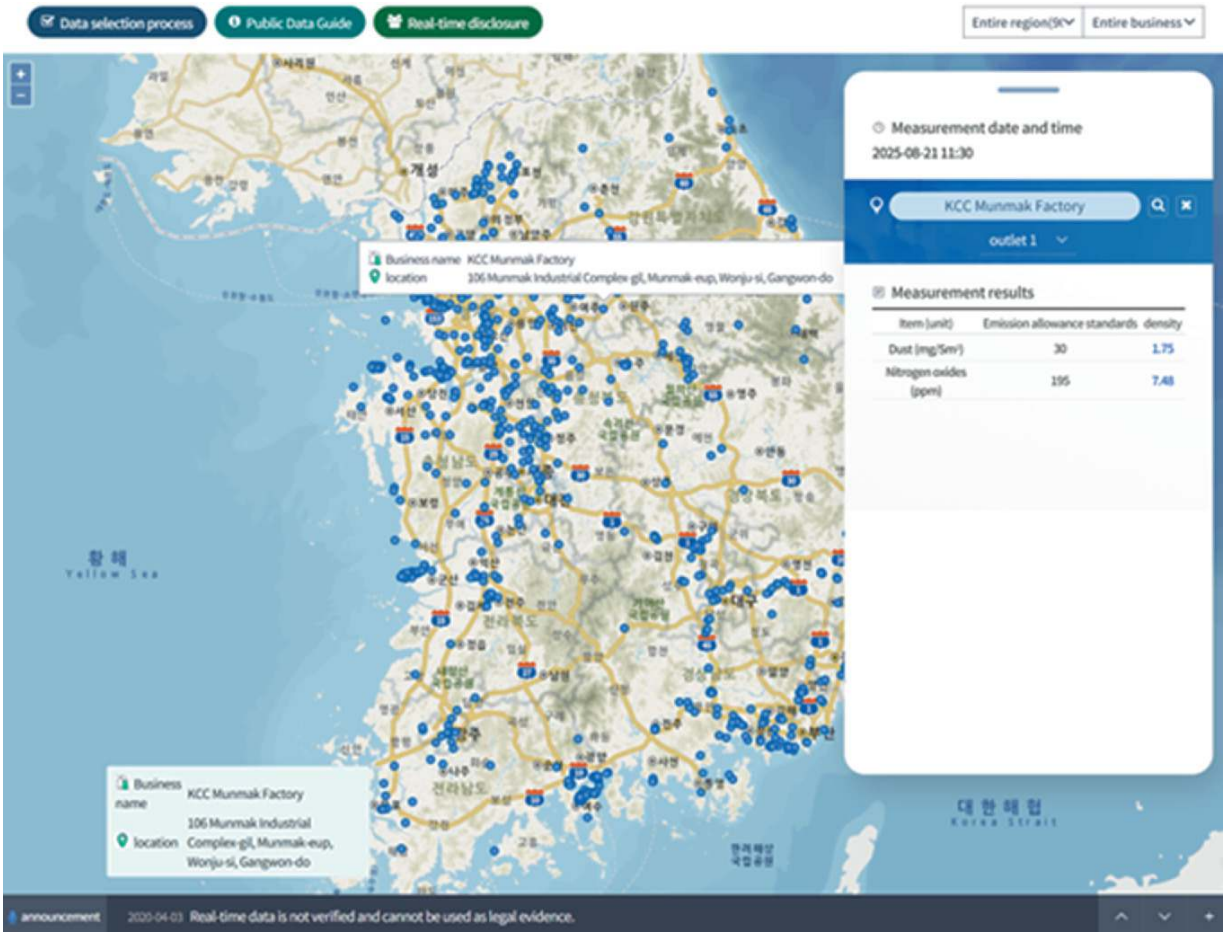


Fig. 7.3 Stack remote monitoring system.
Source Korea Environment Corporation (www.keco.or.kr)

Fourth, the current e-government service represented by Government24 is not a personalized service but a one-size-fits-all service for the general public. When users log into Government24, the same menus and layout are provided regardless of the user. The system neither recognizes the user nor has any information on the user's health, financial, educational, or residential circumstances. It only provides information on processing the civil services they apply for. Consequently, users must identify all relevant information themselves and then find and apply for related services based on that information. This often leads to missing out on benefits provided by the government or failing to meet obligations.

The AI government agent, which serves as the single point of access in AI government, thoroughly addresses the four issues mentioned

above. Unlike Government24 under the current e-government system, the AI government agent provides services in various forms tailored to the characteristics of digital devices—via Internet browser, app, or chatbot—and in any format desired by the user, such as text, voice, sign language, or gestures. In particular, unlike current chatbots that offer short-answer Q&A-based services, this AI-based conversational chatbot can engage in natural, human-like conversations, precisely analyze user requests, and provide accurate responses and processing, which drastically improves service convenience (Fig. 7.4).

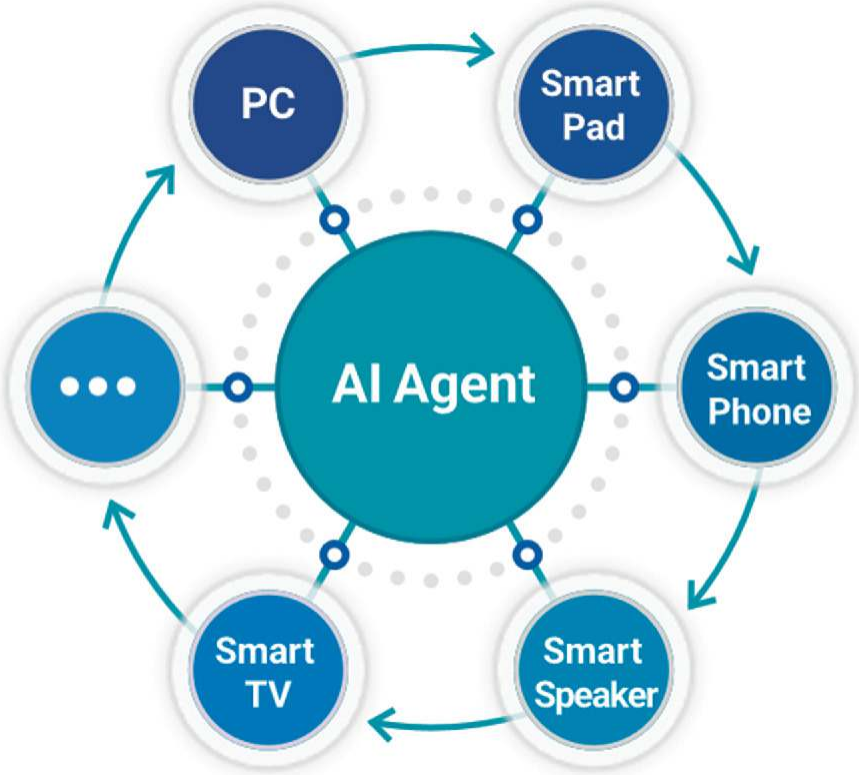


Fig. 7.4 Conceptual diagram of interdevice connectivity for the AI government agent. *Source* Author's compilation

The AI agent is initially installed on one of the user's digital devices. The users can then automatically install and configure the AI government agent on other devices through interaction with the initially installed agent, without having to install and set it up individually on each device. For example, as shown in Fig. 7.5, if the AI government agent is installed on a PC at home, the user can instruct it

by voice to install the AI government agent on other devices connected to the home network—such as a smart pad, smart speaker, smartphone, or smart TV—and the agent handles all tasks automatically. Notably, the AI government Agents installed on each device are not separate agents, but a single AI government agent connected through a network. Therefore, even if a user starts a task with the AI government agent on a PC and switches to a smart pad, the agent on the smart pad is the same as the one on the PC, so the task can be executed seamlessly.

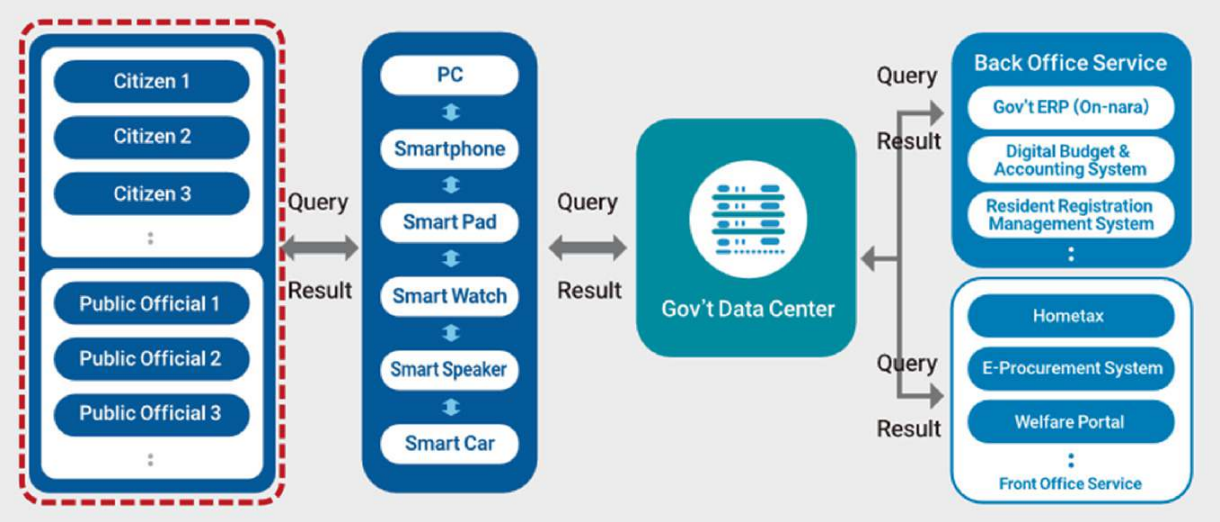


Fig. 7.5 Conceptual diagram of service connectivity of the AI government agent.
Source Author's compilation

The AI government agent provides perfect connectivity to all services built and operated by the government. The current Government24 under e-government only offers civil services to the general public, whereas the AI government agent connects not only to all civil services offered by Government24 but also to all services used by public officials for administrative work, and even to services unknown and inaccessible to individuals because the government has not yet opened them to the public.

Finally, the AI government agent delivers personalized services tailored to an individual's circumstances. It recognizes the user's occupation (e.g., office worker, public official, teacher, soldier) and work status (e.g., active, on leave, parental leave, medical leave), and interacts

in real time to provide and process the services needed. For example, an accounting official can start their day by connecting with the AI government agent to review the progress of tasks from the previous day, set priorities for current tasks, and manage them sequentially. Later, from home, they can use the AI government agent on their PC to check available welfare and education benefits for their 3- and 5-year-old children and submit applications, which the agent processes automatically.

Case Study

Provision of Services by the AI Government Agent

A user accesses the national tax service from home through the AI assistant and checks three received tax bills. The user asks the AI assistant to review the appropriateness of the three taxes. The AI assistant then examines whether the imposed taxes are reasonable. After reviewing the three cases, the AI assistant finds that an excessive tax was imposed on the user's real estate disposed of three months earlier. The user, in consultation with the AI assistant, drafts an objection letter regarding the identified error and submits it to the National Tax Service.

A few days later, traveling in an autonomous connected car, the user becomes curious about the status of the objection and, via voice command, calls the AI assistant and connects to GAIP. The user requests an update on the submitted objection. Recognizing that the user has accessed the system early in the morning via the autonomous connected car, the AI assistant checks GAIP for the status of the submitted objection and verbally reports the progress to the user.

A few days after that, the AI assistant notifies the user watching TV via smartphone that the objection has been processed. The user then connects to the AI assistant through a smart speaker linked to home and reviews the detailed results of the objection on the large TV screen. Upon reviewing the results, the user realizes that during the tax filing process, a donation of 72,000 USD to an overseas university had been

omitted, resulting in an excessive tax charge. The user reports the omitted donation to the National Tax Service through the AI assistant.

After reviewing the newly submitted information, the National Tax Service corrects the excessive charge and issues a revised tax bill, which is delivered to the user via the AI assistant. The user then pays the corrected tax through the AI assistant.

7.3 The Brain of the AI Government: Government AI Platform (GAIP)

Currently, Government24 aims to provide personalized services to each citizen, but in reality, it offers largely standardized services to all users. Regarding the differentiated services provided, the only difference is that individuals can log into their accounts, search for the services they want or are interested in, and check the usage history of the services they have applied for. Specifically, when citizens log into their Government24 account, they are provided with the basic civil services to search for and check the civil services they have applied for so far. In addition, frequently used services and bundled services based on life events—such as moving-in report, happy childbirth, and all-day care—are also provided.

However, as shown in Fig. 7.6, these services are not personalized for "me," rather they are provided to all citizens who fall under such service types. Thus, despite the inconvenience of having to find and apply for each civil service related to moving (e.g., significantly reduced), it still does not provide truly optimized personalized services, and citizens still have to search and apply for the services they need.

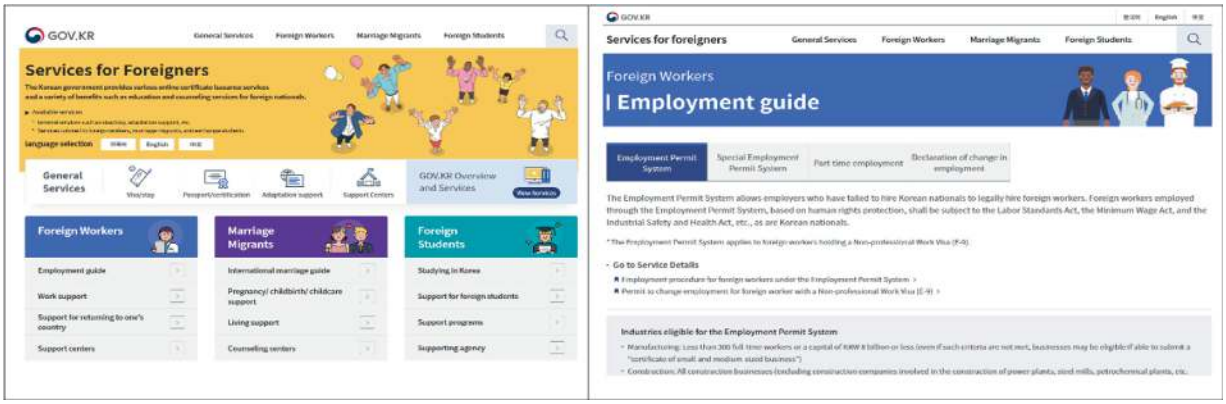


Fig. 7.6 Screenshot of Government24.
Source Government24 (www.gov.kr)

Although Government24 aims for personalized services, why is it unable to provide them as intended and ends up offering standardized services? There are two major reasons. First, all government systems built so far, including Government24, are rule-based systems. A rule-based system operates based on predefined rules and performs a predetermined action when specific conditions are met. These systems are typically expressed in "if–then" structures and are used to systematically organize and apply expert knowledge.

Taking Government24 as an example: when citizens log into their accounts and select "Civil Service → Application & Information" from various menus, the system executes the command defined during development for "Application & Information" and displays the result. In other words, every menu in Government24 has predefined actions at the development stage, and if an undefined menu option is selected or an unsupported request is submitted, the user receives "service not available" or "error" message.

All services provided by the current e-government are based on predefined rules, however, the political, economic, social, and health circumstances of each citizen—and their requests from the government—differ and constantly change. Therefore, it is practically impossible to predefine these individual differences and needs as rules in building a service such as Government24 for all citizens.

Second, detailed data on the environment and needs of each individual citizen are not available. The environment surrounding each citizen constantly changes. For example, if I were to suddenly receive a diagnosis of stomach cancer from a health checkup, my health environment would rapidly change. I would have to visit hospitals regularly, adjust my diet, and exercise properly. This diagnosis would not only affect my health environment but also deteriorate my economic environment, as I would have to take leave from work for treatment. Consequently, my service requests to the government would also change significantly.

Here is another example. If I fall victim to a rental scam and have no funds available, I would need services such as the arrest of the scammer and a loan to secure a place to live. To provide services tailored to such individual situations and needs, detailed data on each individual are essential.

However, services like Government24 are not designed to collect such data. All data are siloed by ministry and system, and access is strictly restricted. As a result, Government24, lacking data on each citizen's situation and needs, can only offer standardized services for citizens to search and apply for on their own.

In contrast, AI government provides services based on AI-based, not rule-based systems. As shown in Table 7.1, rule-based and AI-based systems solve problems in different ways. Rule-based systems operate according to explicitly defined rules, whereas AI-based systems learn from data, adapt, and create or improve rules autonomously.

Table 7.1 Rule-based system vs. AI-based system

This can be understood by imagining how humans solve new problems. When a human faces a new problem, they seek out various information and try to solve the problem using that information in diverse ways. Similarly, when a user submits a request to the AI-based system in AI government, the system searches through government data to find the most effective way to handle that request. That is, AI-

based government services do not rely solely on predefined rules but can generate and optimize new rules autonomously to meet user demands.

In the AI government, providing various "AI-based systems" services requires establishing a GAIP. Then, what is the difference between the Government Artificial Intelligence Platform and government AI-based systems?

Government AI-based systems, as shown in Table 7.2, are single, end-user-focused, specific public services designed to address automated civil service responses, disaster alerts, and traffic congestion forecasting. In contrast, the Government Artificial Intelligence Platform provides the technical infrastructure and environment necessary to develop and operate such systems.

Table 7.2 Government AI-based services vs. Government AI Platform

The Government Artificial Intelligence Platform consists of two components. The first is the AI model, which refers to mathematical and statistical algorithms that analyze data and learn patterns to solve specific tasks or problems—examples include GPT developed by OpenAI, Gemini developed by Google DeepMind, and LLaMA developed by Meta. The AI model is the core component of the AI platform, similar to the brain in the human body. Based on given input data, it performs tasks such as prediction, classification, generation, and inference.

Government AI models are generally developed in two stages. The first stage is the creation of the Government Artificial Intelligence Base Model. As shown in Table 7.3, the base model is a model pre-trained on large-scale datasets and provides a strong foundation that can generally be used for various tasks. This base model can be compared to a person with broad general knowledge, having studied all kinds of information in the world but lacking deep or specialized expertise in specific fields such as medicine, physics, or chemistry.

Table 7.3 Datasets used to train the base model of ChatGPT

The second stage is the creation of the Government Artificial Intelligence Fine-tuning Model. This model is optimized by further training the previously created base model to meet specific tasks or objectives. The fine-tuning model involves refining the generally trained base model to suit specific domains, datasets, or tasks, thereby achieving higher accuracy and relevance. Because the base model is trained on data such as website content, e-books, and more, it is capable of capturing generalized patterns and knowledge. Through fine-tuning, it acquires capabilities tailored to specific sectors or applications. For example, if the base model is intensively trained on data related to defense operations, weapons, missions, and terrain, it becomes a government defense AI fine-tuning model. Likewise, if it is trained on data related to diseases, symptoms, treatments, and prescriptions, it becomes a government healthcare AI fine-tuning model (Fig. 7.7).

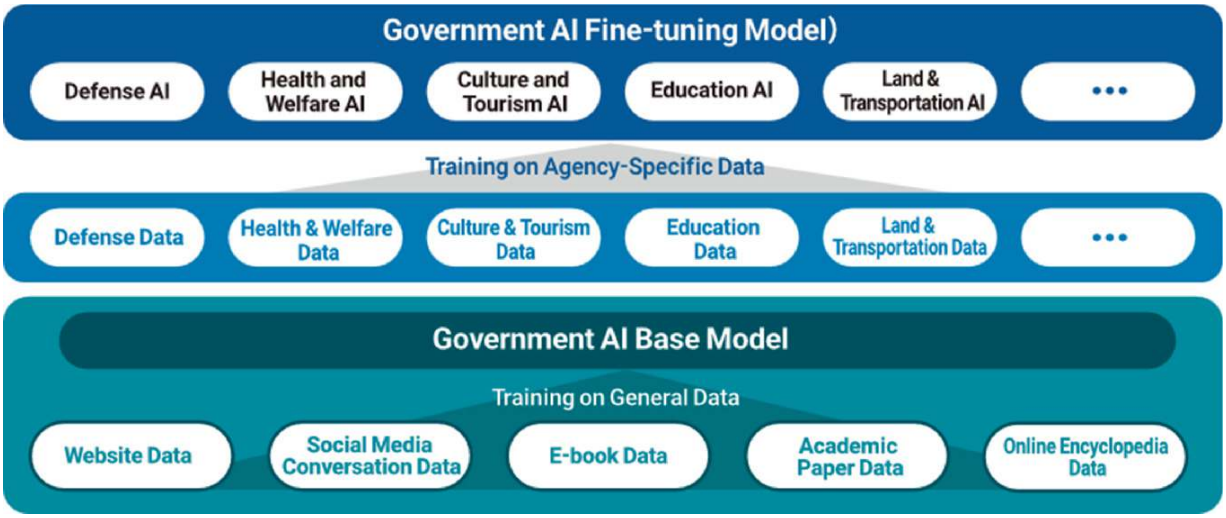


Fig. 7.7 Procedure for creating Government AI models.
Source Compiled by the author

The second component of the Government Artificial Intelligence Platform is data. The data necessary for developing government AI models are categorized into two types. The first type is used to create the Government AI Base Model and consists of diverse, high-quality,

large-scale datasets spanning all fields. These include data from numerous websites, simple notification service (SNS) conversations, e-books, academic papers, encyclopedias, and more.

The second type consists of domain-specific, expert-level data required for creating the Government AI Fine-tuning Model. In the medical field, this would include clinical records, X-ray images, and prescription data. To be usable, such specialized expert data must meet the three conditions shown in Table 7.4.

Table 7.4 Data requirements for creating Government AI fine-tuning models

To improve the performance of the Government AI Fine-tuning Model, all data held by government ministries must be used for training. "All data" include not only data generated by systems but also instructions, regulations, reports, presentations, and statistics created by individual public officials in the course of their duties. Omitting any of these data sources can degrade the performance of the fine-tuned government AI model.

7.4 Search Engine for All Government Services: Whole-of-Governmental Service Catalog

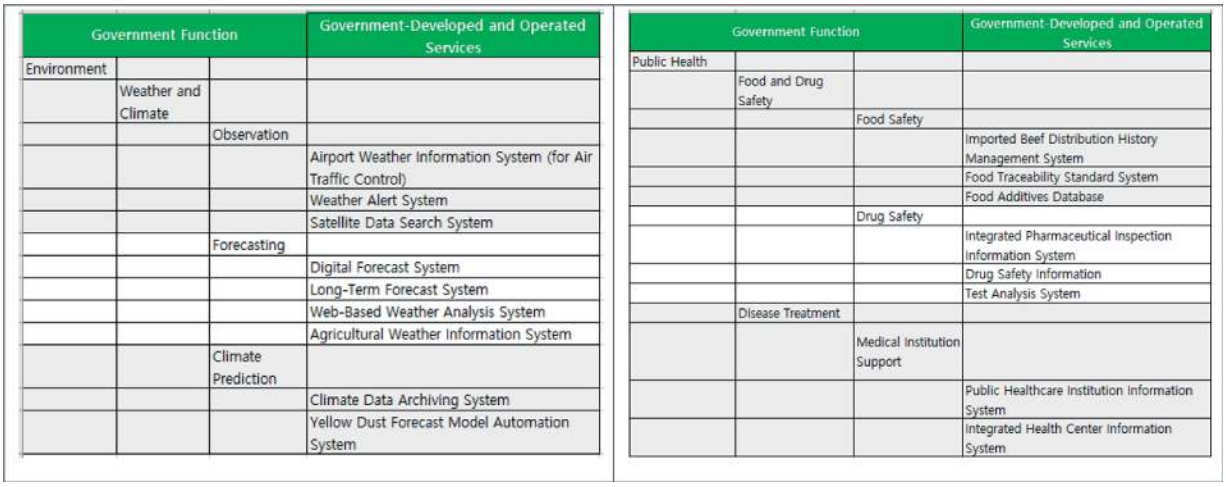
The Pan-Governmental Service Catalog is a database that systematically collects information on services that are being developed and operated by all ministries and agencies of the government. Similar to the product catalog of an online shopping mall, which allows users to view all available items, the Pan-Governmental Service Catalog allows users to view all government services. In other words, the Pan-Governmental Service Catalog can be described as the metadata of government services.

As shown in Table 7.5, the Pan-Governmental Service Catalog consists of the service list, metadata, service classification system, and

service APIs. Among these, the most important component is the service classification system.

Table 7.5 Components of a Whole-of-Government Service Catalog (example)

As shown in Fig. 7.8, the service classification system is a hierarchical framework that organizes services based on various criteria—such as the functions performed by the government, service types, and organizational structure. This system clarifies the relationships between services and provides users with a consistent information structure without duplication, enabling the Government Artificial Intelligence Platform to quickly and easily locate and utilize the necessary services.



Government Function			Government-Developed and Operated Services
Environment			
	Weather and Climate		
		Observation	
			Airport Weather Information System (for Air Traffic Control)
			Weather Alert System
			Satellite Data Search System
		Forecasting	
			Digital Forecast System
			Long-Term Forecast System
			Web-Based Weather Analysis System
			Agricultural Weather Information System
		Climate Prediction	
			Climate Data Archiving System
			Yellow Dust Forecast Model Automation System

Government Function			Government-Developed and Operated Services
Public Health			
	Food and Drug Safety		
		Food Safety	
			Imported Beef Distribution History Management System
			Food Traceability Standard System
			Food Additives Database
		Drug Safety	
			Integrated Pharmaceutical Inspection Information System
			Drug Safety Information
			Test Analysis System
	Disease Treatment		
		Medical Institution Support	
			Public Healthcare Institution Information System
			Integrated Health Center Information System

Fig. 7.8 Service classification by government function (example).
Source Ministry of the Interior and Safety (2024)

7.5 Small, Lightweight, and Modular LEGO-Like Government Services

The characteristics of government services provided under e-Government can be broadly summarized in three aspects. First is the large scale of projects. To provide more convenient services to citizens, fragmented services must be integrated and linked together, which

increases the complexity and difficulty of system development, inevitably leading to large-scale projects costing hundreds of billions of dollars. For example, in April 2023, the Digital Platform Government Committee announced its plan to interconnect and integrate over 17,000 government systems to establish the “DPG Hub,” the top-level integration platform designed to safely connect, integrate, and utilize data and services from public and private sectors. Linking such a large number of systems naturally increases complexity and difficulty, resulting in project scale expansion.

However, large systems built through such large-scale projects suffer from a critical issue: identifying the root cause of a problem becomes extremely difficult. For example, identifying a problem in a system with a single function is much easier than doing so in a system with hundreds of functions. Large-scale systems typically consist of hundreds to thousands of functions, and since they are also interconnected with other large systems for service convenience, pinpointing the source of any issue becomes even more challenging.

A representative example of this challenge is the national administrative network outage in 2023. On the morning of November 17, 2023, a system failure occurred in municipal and provincial administrative networks across the country, rendering administrative welfare centers unable to process civil affairs or issue documents for three days. The outage extended beyond administrative services, causing banking disruptions, suspension of book loans at national libraries, paralysis of tax and real estate services, disruption of Government24, inability to locate 119 emergency callers, malfunctions in the Ministry of Foreign Affairs’ personnel system, disruptions in the local subsidy management system, shutdown of the police cybercrime reporting website, and many other cascading failures. Despite the government's establishment of a multi-agency task force (TF) involving 14 organizations participating to identify and resolve the issue swiftly, the disruption persisted for two to three months.

Second, most e-government services are centrally managed and operated by the Government Integrated Data Center under the National

Information Resources Service (NIRS), which is part of the Ministry of the Interior and Safety. The integrated center promotes operational efficiency and cost reduction through centralized resource management, prevents redundant investment, and enables optimal resource use. With a unified security system and 24/7 monitoring, it protects critical national information from cyberattacks and data breaches, ensuring stable and continuous service through disaster recovery systems and distributed data centers. Currently, the system includes Data Center 1 in Daejeon, Center 2 in Gwangju, and Center 3 in Daegu.

However, the rise of artificial intelligence has exposed the limitations of the centralized government-only data center model. Developing and operating AI services requires vast computational resources. For instance, OpenAI invested approximately 4.3 million USD just for data training in developing ChatGPT, with total costs reaching hundreds of billions of dollars. To operate ChatGPT, which handles millions of user queries daily, thousands of GPU servers are required, incurring costs of several million dollars per day. A single GPU server typically contains 8–16 GPUs, with each NVIDIA H100 GPU priced between 36,000 to 65,000 USD. Therefore, running just one server can cost 288,000 USD, and operating 1000 such servers would require 288 million USD.

Thus, considering cost and operation, it would be inefficient for the government to independently build and operate data centers to provide AI services, as it did under the e-Government model.

Third, the renewal cycle of e-government services is extremely long. On average, next-generation systems are developed every 10–15 years. This cycle is closely tied to the hardware lifecycle. Network equipment, servers, and storage devices begin to fail more frequently after approximately 10 years, making it difficult to maintain high-quality service. Therefore, governments typically undertake full-scale service renewal and hardware replacement considering this timeline.

However, the emergence of AI has significantly accelerated the pace of technological change compared to the traditional e-Government

environment. Previously, it took several years for a single ICT technology to mature. Now, as illustrated in Fig. 7.9, new technologies and innovations are emerging every few months or every 1–2 years at most. Thus, the traditional approach of replacing systems every 10–15 years is no longer valid. However, pursuing large-scale system renewals every 1–2 years, to keep pace with rapid technological change, would impose an unsustainable burden on the government. On the other hand, if the government continues with 10–15 year upgrade cycles due to cost concerns, it may fail to adopt new technologies in a timely manner, inconveniencing citizens. As a result, the government is confronted with a dilemma between cost efficiency and user convenience.



Fig. 7.9 Time taken for online services to reach one million users.
Source Statista (2023)

In summary, the traditional e-government services can be characterized as large, heavy, and difficult to integrate. In contrast, AI government services are small, lightweight, and easily combinable with other services. Why is that so? First, services in AI government are

small in scope. Unlike e-Government systems such as Government24, which are designed to integrate hundreds or thousands of services into a single platform, AI government services are developed for individual use cases only, making their complexity relatively low. This enables faster and easier troubleshooting when issues occur (Fig. 7.10).

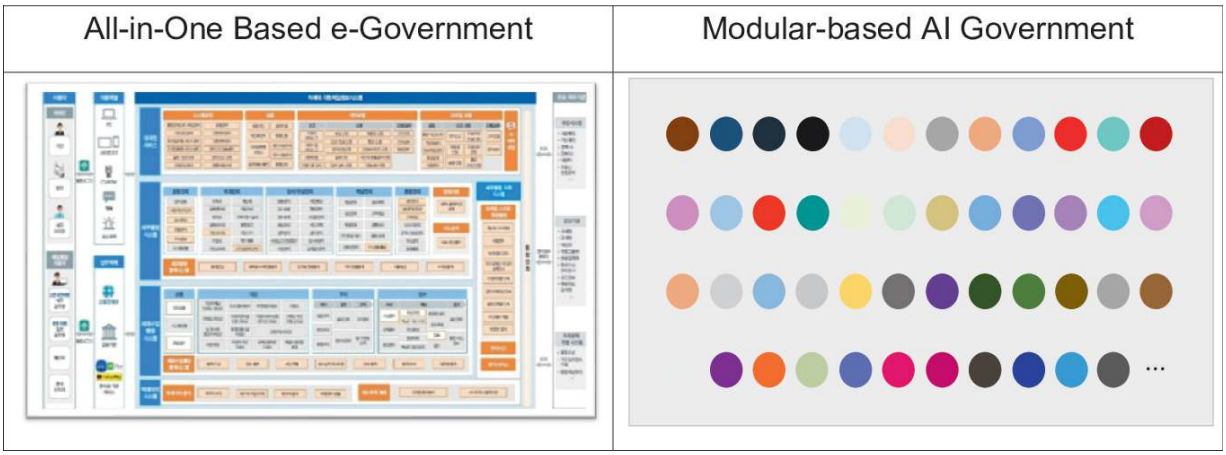


Fig. 7.10 Comparison between waterfall-based e-government and module-based AI government. *Source* Ministry of the Interior and Safety [MOIS] & Korea Local Information Research & Development Institute [KLID], 2024).

Second, AI government services are based on resource-intensive AI technology, making it impractical for the government to build and operate dedicated AI data centers alone. That is, a private cloud environment exclusively designed for AI government services is no longer feasible. However, fully outsourcing government AI services to a public cloud managed by private providers is not an option because of national security and public safety concerns surrounding sectors such as defense, diplomacy, health insurance, and citizen information. Therefore, the government must build a hybrid cloud environment, as shown in Fig. 7.11, to link government-exclusive data centers (private cloud) with commercial cloud providers (public cloud), to ensure service stability and secure sufficient computing resources.



Fig. 7.11 Private cloud environment of e-government vs. hybrid cloud environment of AI government.

Source Sitesbay (n.d.) and Lee (2004)

Third, AI government services are updated frequently on a monthly, quarterly, or yearly basis, depending on the nature of the service. As mentioned earlier, each AI government service is built for a single function, not as part of a bundled system. This makes each service lightweight and simple, allowing easy updates and rapid adoption of new technologies.

As a result, within the AI government framework, numerous microservices are continuously improved and restructured to provide users with more convenient and up-to-date services. This development and operational approach is known as DevOps—a combination of "Development" and "Operations"—which fosters efficient collaboration, automation, and rapid deployment by closely integrating software development and ICT operations (Amazon Web Services, n.d.).

In traditional e-government, development and operations were strictly separated. Large-scale service renewals were costly and typically executed only through long-term plans every 10–15 years, making it difficult to respond quickly to operational issues. DevOps addresses this limitation.

DevOps breaks down large, complex, integrated services into minimal, modular units that are small and lightweight. Each service can easily connect with others via APIs (Red Hat, n.d.). In other words, AI government services are small, lightweight, and modular—like LEGO blocks—allowing flexible combinations with other services.

7.6 High-Quality Machine-Readable Data Based on a Whole-of-Government Classification System

From an AI perspective, three key problems can be identified in the current e-government data environment. First, a whole-of-government data classification system is lacking. As discussed in Part 1, the government's data ecosystem largely comprises three platforms: the Open Government Data Platform, Big Data Platform, and AI Hub. However, a unified data classification system that can interconnect these three platforms has yet to be established. Among all aspects of data, classification is the most critical. Especially today, with the exponentially increasing volume of data and an increase in the proportion of unstructured data, the absence of a clear classification system can lead to hours or even days spent searching and sorting data. In the worst case, invaluable government data may exist but remain undiscoverable amid the sheer volume—much like being unable to locate a specific book in a large bookstore due to a lack of categorization.

Second, although the government has made efforts to collect all the data produced and managed by public institutions through the three platforms, the most crucial data—those produced in the daily workflows of public officials such as policy documents, plans, reports, PowerPoint presentations, and Excel files—are not included. From a data governance perspective, this omission means that a substantial portion of government data is effectively missing from the ecosystem.

Third, there is a shortage of machine-readable data. Recently, the government has made significant efforts to collect machine-readable data through the three platforms mentioned above, and consequently, a considerable proportion of data are now machine-readable. However, the daily work data produced by public officials still mostly exist in non-machine-readable formats on personal computers.

In the AI government, as illustrated in Fig. 7.12, all three data platforms established under e-Government are integrated and managed based on a whole-of-government data classification system to collect 100% high-quality machine-readable data.

No.	Classification Level	Effective Date	Responsible Agency
418	Major Function	20**-03-23	Ministry of Trade, Industry and Energy
419	Major Function	20**-03-23	Ministry of Land, Infrastructure and Transport
420	Major Function	20**-01-27	Ministry of Employment and Labor
421	Major Function	20**-03-20	Ministry of Trade, Industry and Energy
422	Major Function	20**-03-20	Korean Intellectual Property Office
423	Major Function	20**-03-20	Korean Intellectual Property Office
424	Major Function	20**-03-20	Korean Intellectual Property Office
425	Major Function	20**-07-03	Ministry of Employment and Labor
426	Major Function	20**-07-03	Ministry of Employment and Labor
427	Major Function	20**-01-28	Ministry of Employment and Labor
428	Major Function	20**-03-20	Ministry of Environment
429	Major Function	20**-03-20	Saemangeum Development and Investment Agency
430	Major Function	20**-03-19	Ministry of Health and Welfare
431	Major Function	20**-03-23	Ministry of Oceans and Fisheries
432	Major Function	20**-12-14	Ministry of Trade, Industry and Energy / Ministry
433	Major Function	20**-01-24	Korea Customs Service
434	Major Function	20**-06-02	Sewol Ferry Investigation Commission
435	Major Function	20**-06-02	Sewol Ferry Investigation Commission
436	Major Function	20**-03-20	Ministry of Health and Welfare
437	Major Function	20**-07-26	National Fire Agency / Ministry of the Interior and
438	Major Function	20**-07-20	National Fire Agency
439	Major Function	20**-01-27	Korea Fair Trade Commission
440	Major Function	20**-07-26	Ministry of SMEs and Startups
441	Major Function	20**-11-18	Ministry of Personnel Management
442	Major Function	20**-07-26	Ministry of Science and ICT
443	Major Function	20**-12-14	Ministry of Science and ICT

Category	Item Name	Item Definition and Writing Guidelines
Information System Info	Organization Name	Enter the name of the institution that operates and manages the information system and database (DB).
	Information System Name	Enter the name of the information system linked to the target database. (Use the same system name as registered and managed in the Government EA Portal, if applicable.)
	Related Law (Ownership)	Enter the law related to the operation of the information system and data collection.
	Construction Date (Year/Month)	Enter the year and month the information system was constructed.
	Operating Department	Enter the name of the department responsible for operating and managing the information system. (If outsourced, register the actual institution responsible for operating and managing the system.)
	Person in Charge	Enter the name (including position) of the person responsible for operating and managing the information system.
	Telephone Number	Enter the phone number of the person responsible for operating and managing the information system.
	Email	Enter the email address of the person responsible for operating and managing the information system.
	Type of Information System	Select and enter one of the following: development, single-use, or standard system. (Use the same classification as registered in the Government EA Portal (GEAP). If selecting 'single-use' or 'standard', indicate the implementing or user agency as needed.)
	Exclusion from Collection	Indicate with Y/N whether the system is excluded due to being categorized as a system used only for management (e.g., commercial software for security, networks, backups). (If Y is selected, specify the product name or applied task for clarity.)

Fig. 7.12 Example of government function/organization-based data and metadata classification system.

Source Ministry of the Interior and Safety (2024)

In this system, data are first managed according to the government-wide classification framework. This framework categorizes all data produced and managed by the government based on function, organizational structure, metadata, and data types, enabling easier data search and utilization.

In the AI government, all data generated during the daily work of public officials—such as policies, plans, reports, presentations, and spreadsheets—are collected. These collected data are then used in the second phase of the GAIP, as discussed in Sect. 7.3, to fine-tune the government AI model. In this phase, a Foundation Model, pre-trained on large-scale general knowledge, is further trained on real-world data from public officials' daily tasks, which transforms the model into a domain-specific AI with expert knowledge in areas such as administration, diplomacy, defense, and welfare.

Finally, in the AI government, all daily work data produced by public officials are converted into machine-readable formats. To accomplish this, the government is developing a "Whole-of-Government Machine-

Readable Data Converter" to transform previously unstructured, non-machine-readable formats (such as HWP documents) into machine-readable data. Through this process, all government-held data can be continuously integrated into government AI models, maintaining relevance and ensuring high-quality service delivery.

7.7 Problem Solving Through Modular Combination of LEGO-Like AI Government Services

In the e-Government era, when a citizen made a specific request, if a service did not exist for handling that request, the government either developed a new service or had a public official process the request manually. However, as society changes and technology evolves, the types of requests by citizens also change constantly. To address these diverse and rapidly changing demands efficiently, the government has been continually adopting new technologies to develop new services.

Consequently, the number of services and systems the government must manage has grown exponentially, demanding substantial human and financial resources for maintenance. This challenge will continue as long as society evolves and technology advances.

In the AI government, as shown in Fig. 7.13, citizens' ever-changing demands are met not by continuously developing new services, but by combining existing small, lightweight, modular AI government services —structured like LEGO blocks—into various configurations.

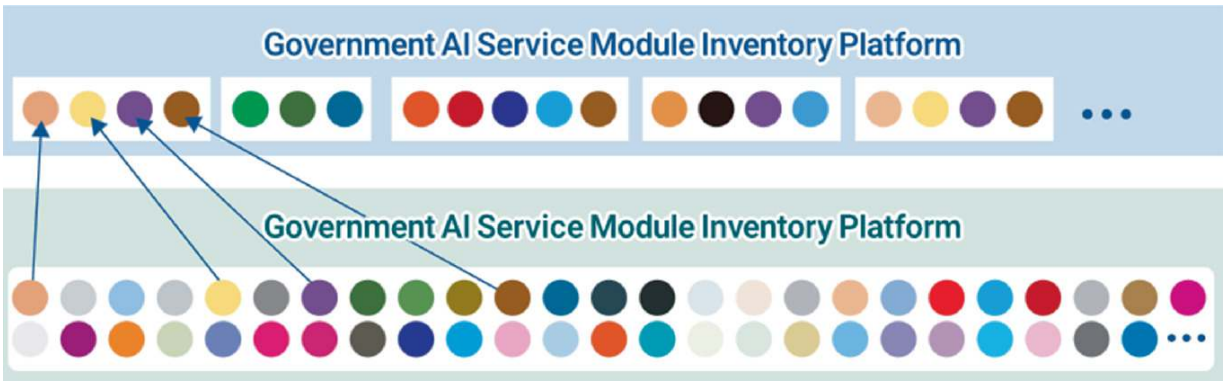


Fig. 7.13 Virtual service created through the combination of modular (legoLEGO-block style) AI government services.

Source Author's compilation

The key point is that when combining these services to create a virtual service tailored to a citizen's request, only the APIs of each service should be used. This is similar to building with LEGO blocks. For example, if constructing the Eiffel Tower with LEGO, you select the most suitable pieces from a large set during assembly.

What matters is that no matter how appealing a particular color or shape may be, if a LEGO block does not physically fit with the others, it cannot be used. Similarly, even if a high-quality service block exists, its code should not be modified to connect with another service. Integration must occur solely through each service's API.

Adhering to this principle ensures that services can be continuously reused and that new service combinations can be created over time.

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2026
S.-J. Eom et al., *Enabling Data-Driven Innovation and AI Government in Korea*
https://doi.org/10.1007/978-981-95-9175-6_8

8. Challenges and Tasks for Implementing the AI Government

Seok-Jin Eom[1] ✉, Yoon-seok Ko[2] ✉, Jeong Eun Park[3] ✉ and Keon Chul Park[4] ✉

(1) Graduate School of Public Administration, Seoul National University, Seoul, Korea (Republic of)
(2) National Information Society Agency, Daegu, Korea (Republic of)
(3) National Information Society Agency, Daegu, Korea (Republic of)
(4) Department of Business Administration, Tech University of Korea, Siheung-si, Korea (Republic of)

✉ **Seok-Jin Eom (Corresponding author)**
Email: sjum21@snu.ac.kr

✉ **Yoon-seok Ko**
Email: ysko@nia.or.kr

✉ **Jeong Eun Park**
Email: jpark0070@gmail.com

✉ **Keon Chul Park**
Email: parkkc07@tukorea.ac.kr

Keywords AI government – Machine-readable data – Modular service design – Data quality assurance – National data classification

8.1 Overview

For the existing e-government system to evolve into an AI government, a fundamental and comprehensive transformation is required. Traditional e-Government systems are rule-based, relying on predefined rules or conditions to perform specific tasks. In such systems, when a value A is input, it is processed according to the developers' designed rules and procedures, producing an expected output B.

Developers verify system integrity by comparing the actual output with the expected result. If the output deviates from the expected value, they review and correct the designed rules and processes to ensure proper functioning.

In contrast, this rule-based methodology alone is no longer sufficient in an AI government. As explained in Chap. 4, similar to how humans acquire knowledge to perform tasks, the government AI model —the core of AI government—learns from government-produced and government-managed data to identify efficient processes and rules. Using these optimized processes and rules, it then assembles small, lightweight, LEGO-block-style services to perform the assigned tasks.

In e-Government, humans designed and developed every system component, giving them complete control over its operation. In AI government, however, the government AI model designs the most efficient processes and rules for a given task on behalf of humans and combines LEGO-block-style government services to create virtual services that can be dynamically generated and terminated. Therefore, human intervention at appropriate stages is essential to verify the integrity and reliability of AI-handled tasks. This chapter outlines six key points detailing how and what to verify to ensure the integrity of AI government.

The first is the verification of the government AI model, which forms the foundation of the government AI platform. Analogous to the human brain, it is essential to identify and assess any flaws or limitations in the model.

Second is converting all data produced and managed by the government into machine-readable formats. Existing data that are not

machine-readable should be promptly converted and fed into the government AI model for learning. Newly generated data should also be machine-readable to enable the model to learn as efficiently as possible.

Third is establishing a nationwide data classification system. With the exponential growth of data resources available at the national level, the volume of data the government AI model can learn from is increasing rapidly. However, this growth also drives up the costs of data collection, preprocessing, usage, and management. A standardized data classification system is therefore necessary to manage these expanding resources efficiently.

Fourth is verifying the quality of data supplied to the government AI model. Because AI is highly sensitive to its input, the data provided are the primary means by which humans can influence model performance. Data quality is the most critical factor affecting the government AI model's effectiveness, making national-level verification of data quality essential.

Fifth is transforming large, bundled services into small, lightweight, and freely combinable LEGO-block-style services. For the government AI platform to operate effectively, all services should follow this modular design, allowing the AI to complete tasks by dynamically combining these services as needed.

Sixth is verifying virtual services generated through the combination of modules to perform assigned tasks. While the government AI platform can rapidly process tasks by automatically combining LEGO-block-style services, the integrity of the resulting virtual services must be checked to prevent errors and potential issues.

8.2 Verification of Government AI Models

The complexity of artificial neural network-based AI models makes it extremely difficult for humans to accurately predict their outcomes or fully understand the internal processes that generate those outcomes. Therefore, the government must continuously verify that the government AI models—the foundation of the AI government—are

error-free and monitor for issues that may arise when models are connected or integrated.

The AI government comprises a core model overseeing all government services (similar to ChatGPT), domain-specific models for sectors such as healthcare, transportation, and agriculture, and individual service-level models. All of these models are built on artificial neural networks. However, they can occasionally generate inaccurate or inappropriate responses to queries. A fundamental challenge is that, due to the inherent complexity of neural networks, it is extremely difficult to identify which combinations of neural nodes lead to incorrect outputs.

This marks the beginning of the government's challenge in adopting AI. Government operations and public administrative services must be grounded in integrity. For example, if financial aid intended for low-income individuals is mistakenly distributed to high-income earners, the resulting confusion and conflict could severely undermine public trust and disrupt state functions. Therefore, the government must continuously examine all AI models to minimize errors and ensure that failures do not occur when models are interconnected or combined.

Rigorous verification of such AI models requires interdisciplinary collaboration. Experts from all relevant fields—mathematics, medicine, biology, linguistics, engineering, and others—must work closely together. In this context, establishing a "National AI Model Verification Center" should be considered as a policy option to facilitate continuous and collaborative oversight.

8.3 Transition to Machine-Readable Data

The primary reason for adopting artificial intelligence is its ability to expand the range of tasks it can perform through continuous learning from new data. In other words, if an AI system can handle 10 tasks in its first year, it can learn additional data in its second year to handle 20 tasks, and by the third year, further expand to 30 tasks—without

developing new systems for each task, instead relying on ongoing data-driven learning.

However, a crucial prerequisite is that newly generated public sector data must be continuously and seamlessly supplied to the government AI model in a machine-readable format. Regardless of its value, data that are not machine-readable will be difficult—or impossible—for the AI model to learn from. This is analogous to speaking Korean to someone who only understands English. If non-machine-readable data are provided, they must either be converted—which incurs substantial computing costs, including server and storage resources—or discarded as unusable. Data that are discarded cannot contribute to learning, ultimately reducing the platform's performance. Therefore, all data supplied to the government AI platform must be fully converted into a machine-readable format before ingestion.

Machine-readable data refer to data that can be read, edited, transformed, and shared across software systems without being limited to a specific application. Representative formats include CSV, JSON, XML, RDF, and LOD.

Currently, a significant portion of government-generated data are not machine-readable. For example, many government reports are produced in PDF format, which is often not easily editable or transformable and can only be opened with specific software—making it a prime example of non-machine-readable data. Similarly, reports created in Microsoft Word can only be read, modified, or converted using that specific software, which disqualifies the data from being considered fully machine-readable.

Therefore, how can we efficiently convert vast amounts of data into a machine-readable format for use by the government AI platform? The answer lies in integrating a machine-readable data converter at the data creation stage, to ensure that all documents and related content are generated in machine-readable formats from the outset. A key function of the converter is its ability to transform even unstructured data—such as text, images, audio, and video—within a document into

machine-readable data that can be learned by the government AI platform.

For example, suppose a civil servant writes a flood damage report in Microsoft Word. To provide a comprehensive overview, the report would include not only numerical data but also photographs of the damage. In this case, text and numbers are relatively easy to convert into machine-readable formats, but image data require data labeling to help the AI system understand what is depicted in the photos. Thus, the machine-readable data converter should ensure that all data are transformed in a way that allows the government AI platform to "see" and understand the content just as a human would.

8.4 Establishing a National-Level Data Classification Model

In the initial stages of implementing the government AI platform, the most important factor is how quickly abundant data can be supplied to the government AI model for training. In general, the more data the model learns from, the more tasks it can potentially handle, and the increase in the platform's task-processing capabilities ultimately leads to improved government efficiency. However, the data used to train the model are not exclusively created for learning purposes—rather, they are generated for a wide range of purposes beyond just training.

Therefore, the government must manage the massive amounts of data used in model training in a way that enables easy search, use, sharing, and open access. This responsibility extends not only to data used for training but also to all the data the government possesses. Managing this vast array of data effectively requires establishing a national-level data classification system.

To this end, two main considerations must be addressed. The first is a metadata classification system. When discussing data classification, most people refer to metadata classification systems, as these are relatively well-known. Future cross-border data exchanges necessitate that the metadata classification system align with global standards. As

of 2024, the most widely used standard for metadata classification is DCAT, maintained by the *World Wide Web Consortium* (W3C). DCAT defines how a government's metadata classification structure is expressed.

By adopting DCAT, the government can not only make its data more searchable, but also access datasets from other countries that have adopted the same standard, thereby promoting cross-national data sharing and reuse. Currently, most countries are building their national metadata classification systems based on DCAT. Accordingly, any government preparing to build or update its metadata system is strongly encouraged to adopt a DCAT-based approach to enable future global interoperability.

Figure 8.1 shows a metadata classification system established using DCAT, based on ten metadata items commonly used by ASEAN Member States: title, publisher, updates, data portal, URL, file type, description, download URL, file name, and license. Although metadata classification systems that categorize data based on attribute information are highly effective for managing and retrieving data quickly, recent trends in big data and artificial intelligence have led to more complex data usage. Data are no longer confined to a single service domain but are instead leveraged across service delivery, analysis, sharing, and open access.

```
@prefix dcat: <http://www.w3.org/ns/dcat#> .
@prefix dct: <http://purl.org/dc/terms/> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .

<http://example.org/dataset/1> a dcat:Dataset ;
    dct:title "Title" ;
    dct:publisher [ a foaf:Organization ; foaf:name "Publisher" ] ;
    dct:issued "2024-06-29"^^xsd:date ;
    dct:modified "2024-06-30"^^xsd:date ;
    dcat:contactPoint <http://example.org/dataportal> ;
    dcat:landingPage <http://example.org/dataset/1> ;
    dct:format "File Type" ;
    dct:description "Description" ;
    dcat:distribution [
        a dcat:Distribution ;
        dcat:downloadURL <http://example.org/download/file1> ;
        dcat:mediaType "application/zip" ;
        dct:title "File Name" ;
        dct:license <http://example.org/license>
    ] .
```

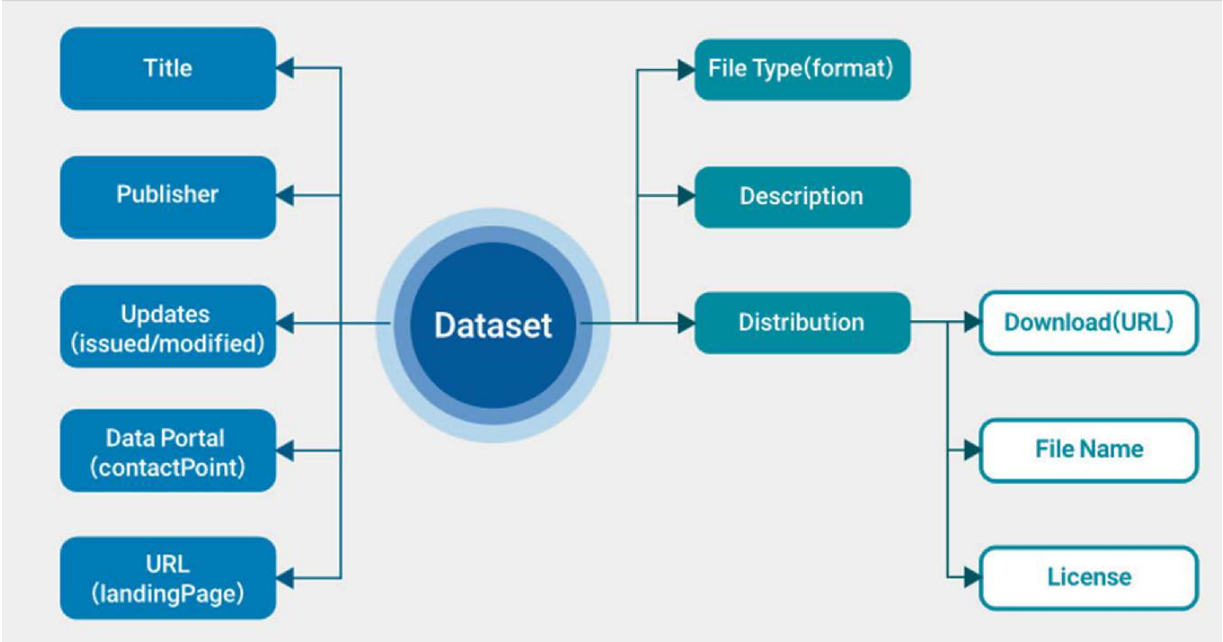


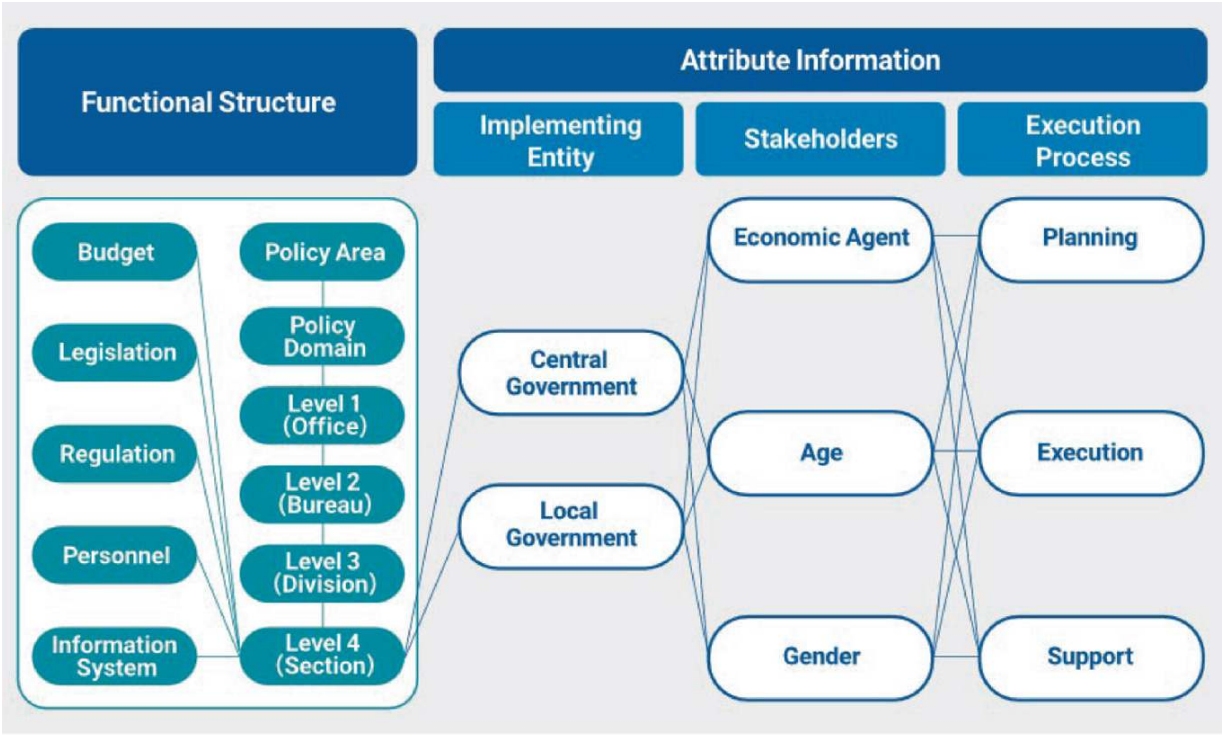
Fig. 8.1 Example of a metadata classification system using DCAT.
Source Author's compilation

Thus, in addition to classifying data by metadata attributes, governments now require a more holistic classification system that allows a comprehensive view of all government-owned data—along

with relationships to organizations, functions, systems, and other surrounding elements. One of the most effective approaches is to classify data based on government functions and organizational structures.

This is especially important because, although government organizations often change to meet administrative needs, the functions they perform remain relatively stable. Therefore, one of the most effective methods for classifying and managing data—and their contextual relationships—is to base the classification on governmental functions.

To build the function-based classification system illustrated in Fig. 8.2, the first step is to identify all the functions the government performs. For each function, the system should log all related information, such as responsible entities and procedures, stakeholders, legal frameworks, budgets, information systems, and associated datasets. Similar functions should be linked to one another. This function-centered classification system enables flexible and multifaceted data searches, making it easier to understand the interconnections between datasets, which greatly enhances data utility and governance.



Common Gov't Biz Functions for Health	Code
Health Policy and Planning	01
Health Service Delivery	02
Disease Control and Prevention	03
Health Regulation	04
Emergency Preparedness and Response	05
Health Information Management	06
Healthcare Financing	07
Health Research and Development	08
Human Resource Management	09
Health Education	10
International Health Cooperation	11

Common Gov't Biz Functions for Agriculture	Code
Agriculture Policy and Planning	01
Production Support and Services	02
Plant Health and Quarantine	03
Food Safety and Quality Assurance	04
Livestock and Veterinary Services	05
Agriculture Regulation	06
Disaster Mangement and Emergency Response	07
Agriculture Research and Development	08
Agriculture Information System Management	09
Agriculture Financing	10
Agriculture Trade and Marketing	11
Human Resource Development	12
Agriculture Education	13
International Agriculture Cooperation	14

Fig. 8.2 Concept and example of function-based data classification system.
Source Author's compilation

In addition to the two classification systems mentioned above—metadata- and function-based—governments must also explore other tailored approaches to effectively manage vast volumes of data. Once the transition to AI government is complete, data will become the most critical factor in determining the performance of government AI

models. Therefore, managing data should be treated as one of the highest policy priorities.

8.5 National Data Quality Verification for Ensuring Data Integrity

The most significant factor influencing AI performance is data. Most problems related to AI originate from the training data. For example, Amazon developed an AI-based recruiting tool to select the most suitable new employees for the company. This tool learned patterns from resumes submitted to Amazon over the past ten years and evaluated applicants accordingly. However, over the past decade, the ratio of male to female hires at Amazon was 6:4 in favor of men. As a result, the AI-based recruiting tool, having learned from this male-biased data, rejected resumes that included the word “woman” and only recommended male candidates with over ten years of experience. Ultimately, Amazon scrapped this AI-based recruiting tool in 2018, despite having invested a large budget in its development since 2014. This case clearly illustrates that the quality of data used to train AI can have a decisive impact on its performance.

In government AI models, data integrity is even more critical. If the data learned by a government AI model contain errors or biases, the services provided to the public can also be biased or erroneous, potentially causing significant harm to citizens. Moreover, the government can lose public trust, which would severely affect its ability to function effectively. For instance, if a government AI model fails to identify vulnerable groups such as persons with disabilities or basic livelihood security recipients due to flawed training data, and those individuals are not recognized as socially disadvantaged, they may be deprived of benefits and suffer greatly. This could intensify unnecessary social controversies and further erode public trust in the government. Another example would be a government AI platform used in public servant recruitment learning from biased data, similar to the Amazon case, and ending up favoring specific groups. This would also become a

serious social issue, sparking public debate. Therefore, only data subjected to certified government quality assurance processes and meeting integrity standards should be used for the training of government AI models.

To ensure the integrity of data used for training government AI models, two key aspects must be considered. First, how to measure data quality to guarantee its integrity? The numerous indicators for assessing data quality include accuracy, completeness, consistency, timeliness, validity, conformity, relevance, and uniqueness. Ideally, data should be verified against all indicators, but considering cost and time constraints, this is practically impossible. Only the truly necessary indicators should be selected and measured. Hence, we recommend four quality indicators—diversity, syntactic accuracy, semantic accuracy, and validity—based on Korea's "AI Training Data Construction Project" initiated in 2020.

Since 2020, the Korean government has invested 1 billion USD in over 860 AI training data construction projects. The numerous trials and errors in these efforts, have empirically confirmed that ensuring the integrity of the data used to train AI requires verifying four indicators: diversity of the data; accuracy of its structure and format; precision and accuracy of labeled data; and validity, which reflects improved AI model performance after training. Therefore, the four quality indicators outlined in Table 8.1 must be satisfied to ensure the quality of the data used in government AI models.

Table 8.1 Quality measurement indicators for data used in training government AI models

Second, determining who will verify the data used for training government AI models is crucial. As mentioned earlier, the performance of a government AI model depends significantly on the quality of data it learns from, and training on flawed data can cause significant societal disruption. Given the explosive growth in data expected for AI training in the future, the government must seriously consider how to verify

such data. Table 8.2 presents examples of detailed indicators that can be used to measure data quality across different categories.

Table 8.2 Detailed indicators by data quality measurement category (examples)

Third, a dedicated organization must be established to verify the vast amount of data that government AI models will learn from. This organization should consist of experts specializing in different data types, such as text, image, audio, video. These experts must establish verification criteria for each of the four indicators defined earlier, depending on the data type. Crucially, as the specific verification indicators applied may vary for each data type, the experts must judge which indicators are appropriate for which data.

Once detailed indicators for measuring data quality by type are designed, as shown in Fig. 8.3, a decision must be made on how to measure data quality. First, one must decide whether the verification will be done manually by humans or automatically by software and second, whether the entire dataset should be verified or only a sample will be extracted for verification. A widely used method involves verifying the entire dataset using automated tools and then having humans manually verify a subset of that data. In this case, it is essential to determine how large the extracted sample should be for human re-verification. Given that this data will be used to train government AI models, as much data as possible should be re-verified by humans to maximize human supervision. Particularly, for critical datasets such as medical data that require 100% quality assurance, human verification must be mandatory.

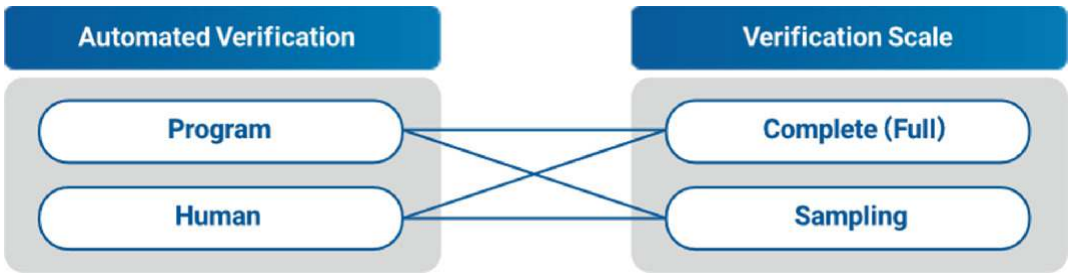


Fig. 8.3 Data quality verification methods.
Source Author's compilation

In addition, the dedicated organization responsible for verification must monitor the entire lifecycle of the data used in government AI model training and perform numerous tasks to ensure that the quality of each dataset is 100% guaranteed.

8.6 Transition to a LEGO-Like Modular System

One of the key advantages of operating a government AI platform is its ability to virtually build new services by combining existing ones to perform assigned tasks. This is similar to how DeepMind's AlphaGo selected the best position to place a stone on the Go board by evaluating countless possible locations. In the same way, the government AI platform tests various combinations of already-developed services and selects the most efficient combination to create a virtual service to complete a given task.

For the government AI platform to freely combine a vast number of pre-developed services as required, all government systems should be redesigned using small, lightweight, and modular LEGO-block-style services. The core requirement is that each service must be independently operable, meaning it should be able to carry out its task autonomously without relying on other systems or services. For example, in an email system, services like "Inbox," "Sent Mail," and "Trash" operate independently and perform their respective tasks without being affected by other components.

Therefore, the government must deconstruct all legacy systems—originally built from interlinked and interdependent services and functions—into the smallest units capable of operating independently. These services must be converted into standalone units that do not rely on others. A crucial requirement is that every service communicates externally only through APIs. Enforcing API-only communication ensures that each service remains fully independent.

However, converting a large monolithic system into the smallest possible independently operating LEGO-block style services is extremely challenging. First, this requires a careful design process to determine how to disaggregate the large system into modular components. Ideally, the original monolithic service should be broken down and redeveloped into LEGO-block services. However, excessive fragmentation can result in a situation where too many small services must be combined to complete even a single task. Such complexity becomes a major cause of service delays and system failures.

Therefore, dividing services into appropriately sized units, balancing granularity, and efficiency are crucial. Determining the right size and scope for each service will vary by use case, requiring the active participation of domain experts familiar with both the business processes and system architecture.

The next priority is the ongoing management of each LEGO-block service. Since each of these services operates independently, they must be regularly monitored and updated to integrate new technologies. To enable this, the existing large-scale maintenance and support teams must be reorganized into agile support units optimized for LEGO-block services, thus enabling a rapid maintenance and upgrade system.

8.7 Validation of Virtual Service Appropriateness

In Sect. 8.6, a single large-scale service was broken down into numerous LEGO-block-style services, and the government AI platform executed tasks by creating virtual services through optimal combinations of these units, each representing different government functions. However, because the newly created virtual service is generated entirely by the government AI platform without human intervention, it is essential for humans supervisors to verify that the virtual service is properly constructed and that its outcomes are accurate.

To ensure accuracy, the composition of LEGO-block-style services, along with the associated processes, data, and outcomes, must be thoroughly verified. As emphasized throughout this report, the government AI platform operates probabilistically. It learns from vast amounts of data and provides the most likely response to a given query, meaning errors can always occur. Therefore, it is critical for the government to identify where human intervention is necessary during platform operations and to ensure such intervention occurs at key points to proactively prevent errors.

References

Agrawal, A., Gans, J., & Goldfarb, A. (2022). *Prediction machines, updated and expanded: The simple economics of artificial intelligence*. Harvard Business Press.

AI Hub. (2023). *2024 AI Hub safe zone operation and management: Request for proposal*. AI Hub.

AI Hub. (n.d.). *안심존이란? [What is the Safety Zone?]*. https://aihub.or.kr/intrcn/safetyzoneintrcn.do

Amazon Web Services. (n.d.). *What is DevOps?*. https://aws.amazon.com/devops/what-is-devops/

Berman, P. (2017). Hacking ideology: pol.is and v Taiwan. *Hacktivism*. Democracy Earth Foundation. https://words.democracy.earth/hacking-ideology-pol-is-and-vtaiwan-570d36442ee5

Bi, X., Chen, D., Chen, G., Chen, S., Dai, D., Deng, C., Ding, H., Dong, K., Du, Q., Fu, Z., Gao, H., & Zou, Y. (2024). DeepSeek LLM: Scaling open-source language models with longtermism. arXiv:2401.02954. https://arxiv.org/abs/2401.02954

BigScience Workshop, Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., & Bari, M. S. (2022). BLOOM: A 176B-parameter open-access multilingual language model. arXiv:2211.05100. https://arxiv.org/abs/2211.05100

Bloomberg. (2023). Introducing BloombergGPT: Bloomberg's 50-billion parameter large language model, purpose-built from scratch for finance. *Bloomberg*. https://www.bloomberg.com/company/press/bloomberggpt-50-billion-parameter-llm-tuned-finance/

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901. https://doi.org/10.48550/arXiv.2005.14165

Choi, Y. S. (2015). Digital healthcare and precision medicine. *The Young Doctor*. https://www.yoonsupchoi.com/2015/12/13/digital_health_precision_medicine/

Cole, S., & Eda, K. (2024). What is artificial intelligence (AI)? *IBM*. https://www.ibm.com/think/topics/artificial-intelligence

Corby, R. (2020). Turing-NLG: A 17-billion-parameter language model. *Microsoft Research Blog*. https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4171–4186). https://doi.org/10.18653/v1/N19-1423

Eggers, W. D., Schatsky, D., & Viechnicki, P. (2017). *AI-augmented government: Using cognitive technologies to redesign public sector work*. Deloitte Center for Government Insights.

Eom, S. J. (2021). Democracy and public administration in the age of artificial intelligence: Focusing on the case of 「e-People」 system advancement. *Korean Public Administration Review, 30*(2), 35–64.

Eom, S. J., & Kim, S. (2024). The multi-dimensional paradox of AI adoption in the public sector: The Korean experience. In *Proceedings of the 25th Annual International Conference on Digital Government Research* (pp. 133–145).

Eom, S. J., Kim, H., Park, J. H., Kwon, H. J., Kwon, H. Y., Lee, S. Y., Park, S. A., Kim, B. H., Kim, S. H., Na, J. M., Ko, G. G., Hwang, J. S., Lee, K. H., Kim, S. J., Kim, D. W., Kim, B. J., Choi, M. S., Chung, Y. J., Lim, D. B., & Hong, J. H. (2021). *AI and future administration [AI와 미래행정]*. Pakyoungsa.

Eom, S. J., Kim, J. H., Choi, N. B., Hwang, H. C., Chung, H. J., & Seo, K. R. (2020). *Digital innovation of government [정부의 디지털 혁신]*. Moonwoo Publishing.

Eun, J. H., Kim, K. G., & Lee, S. W. (2020). A study on strategies for utilizing artificial intelligence for successful conflict management. *Korean Journal of Comparative Government, 24*(2), 193–222.

Fountain, J. E. (2001). *Building the virtual state: Information technology and institutional change*. Brookings Institution Press.

Frost & Sullivan. (2021). *The future of privacy in a hyper-connected world*. Frost & Sullivan. https://ww2.frost.com/frost-perspectives/blog/the-future-of-privacy-in-a-hyper-connected-world/

Gartner. (2022). *Why adaptive AI should matter to your business*. https://www.gartner.com/en/articles/why-adaptive-ai-should-matter-to-your-business

Géron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Google Developers. (n.d.). *Static vs. dynamic inference*. Google. https://developers.google.com/machine-learning/crash-course/production-ml-systems/static-vs-dynamic-inference?hl=en

Han, E. H., Chung, S. W., Kim, J. S., Ryu, J. H., Lim, S. H., Kim, J. S., Woo, Y. S., & Kim, I. B. (2021). Study on improving the operational system of the e-Child Happiness Support Project in towns and villages. *Korea Social Security Information Service*.

Han, E. H., Park, G. B., Choi, B. G., Ahn, Y. J., & Kim, S. (2022). A foundational study for preparing the mid- to long-term development plan of the e-Child Happiness Support System. *Korea Social Security Information Service*.

IBM. (2023). *AI healthcare benefits*. https://www.ibm.com/think/insights/ai-healthcare-benefits

IBM. (n.d.). *What is an AI model?* https://www.ibm.com/think/topics/ai-model

Im, J. W. (2024). What is artificial general intelligence (AGI)? Concept, applied technologies, and impacts on human life. *Samsung SDS*. https://www.samsungsds.com/kr/insights/artificial_general_intelligence_20240417.html

Jeon, C. B. (2021). Why Amazon's hiring AI favored men. *The Korea Daily*. https://www.hankookilbo.com/News/Read/A2021101409500001667

Jeong, J. H. (2020). Operational status and improvement tasks of the big data platform. *Ministry of Science and ICT*.

Jung, D. E. (2024). KakaoBank applies 'XAI' to fraud detection for enhanced security. *The Electronic Times*. https://www.etnews.com/20240628000278

Kim, J. S., Han, E. H., Kim, J. S., Woo, Y. S., & Chung, S. W. (2023). A study on the work experiences of public officials participating in the e-Child Happiness Support Project. *Korean Journal of Social Welfare Administration, 25*(2), 65–99.

Kim, K. J. (2017). *Cases of utilizing IBM Watson in the public sector*. Paper presented at the Policy & Knowledge Forum, Graduate School of Public Administration, Seoul National University.

Ko, Y. S. (2021). Introduction to the big data platform and center construction project. *National Information Society Agency*.

Korea Data Agency. (2021). *2020 data industry status survey report*. Korea Data Agency Research Report.

Korea Environment Corporation. (n.d.). *Allbaro system (national waste management system)*. https://cleansys.or.kr/index.do

Korea Environment Corporation. (n.d.). *Korea Environment Corporation*. https://www.keco.or.kr

Korea Information Society Development Institute. (2021). *Study on diagnosing and enhancing the competitiveness of the domestic data industry*. Korea Information Society Development Institute.

Korea Ministry of Government Legislation. (2025). *Guidelines for standardization of public institution databases*. Korea Ministry of Government Legislation.

Lee, D. H., Heo, J., & Kim, J. M. (2018). Prospects for future jobs in promising software fields. *Software Policy & Research Institute*.

Lee, H. J. (2004). What is a private cloud? Comparison of public and private clouds. *Brainz*. https://www.brainz.co.kr/tech-story/view/id/306#u

Lee, J. S. (2016a). Development process and understanding of deep neural networks: Information and communication. *The Journal of the Korean Institute of Communication and Information Sciences, 33*(10), 102.

Lee, K. S., & Kim, S. O. (2021). Analysis of barriers to AI adoption and diffusion and policy implications. *Korea Information Society Development Institute*.

Lee, K. W. (2016). *Intro to machine learning & deep learning*. National Information Society Agency internal training material.

Lipsky, M. (2010). *Street-level bureaucracy: Dilemmas of the individual in public service*. Russell Sage Foundation.

Margetts, H., & Dorobantu, C. (2019). Rethink government with AI. *Nature, 568*(7751), 163–165. https://doi.org/10.1038/d41586-019-01099-5

Margetts, H., John, P., Hale, S., & Yasseri, T. (2015). *Political turbulence: How social media shape collective action*. Princeton University Press.

Mazzucato, M. (2011). The entrepreneurial state: Debunking public vs. private sector myths. *Journal of Entrepreneurship and Public Policy, 4*(3), 392–394.

Meta AI. (2023). *LLaMA 2: Open foundation and fine-tuned chat models*. https://ai.meta.com/research/publications/llama-2-open-foundation-and-fine-tuned-chat-models/

Meta AI, Mistral AI, EleutherAI, BigScience, Technology Innovation Institute (TII), Databricks, & Together Computing. (2023–2024). *Open source large language models overview: LLaMA 2, Mistral, GPT-NeoX, BLOOM, OPT, Falcon, Dolly, and RedPajama*. Official documents and research presentations.

Ministry of Health and Welfare. (2019). *Performance results of data-based administration: Internal report*. Ministry of Health and Welfare.

Ministry of Science and ICT. (2019). *2018 survey on the status of the data industry*. Ministry of Science and ICT.

Ministry of Science and ICT. (2021). *Fostering the entire data industry! Enacting the 'Basic Data Act' to open the data economy*. Ministry of Science and ICT Press Release.

Ministry of Science and ICT & National Information Society Agency. (2023). *Guidelines for quality management and construction of artificial intelligence training data v3.0*. Ministry of Science and ICT.

Ministry of Science, ICT and Future Planning. (2013). *Plan for fostering Internet new industry promotion plan: Realizing the Internet where ideas change the world*. https://www.korea.kr/archive/expDocView.do?docId=34029

Ministry of SMEs and Startups. (2021). *Analysis of changes in the Korean startup ecosystem*. Ministry of SMEs and Startups.

Ministry of the Interior and Safety. (2021). *Practical manual for the provision and management of public data*. Ministry of the Interior and Safety.

Ministry of the Interior and Safety. (2022). *Digital platform government promotion plan: Public AI services and platform development directions*. Ministry of the Interior and Safety.

Ministry of the Interior and Safety. (2024). *Government function classification system*. Ministry of

the Interior and Safety.

Ministry of the Interior and Safety. (n.d.). *Government24 – Government service, civil complaints, and policy information portal*. https://www.gov.kr

MIT Technology Review. (2021). *The progress issue*. *MIT Technology Review, 124*(4). https://www.technologyreview.com/magazines/the-progress-issue/

Monica. (n.d.). *Monica – Your personal AI assistant*. https://monica.im/

Morris, M. R., Sohl-Dickstein, J., Fiedel, N., Warkentin, T., Dafoe, A., Faust, A., Farabet, C., & Legg, S. (2024, July). Position: Levels of AGI for operationalizing progress on the path to AGI. In *Forty-first International Conference on Machine Learning*.

National Information Society Agency. (2011). *New value creation engine: New possibilities and strategies for big data*. National Information Society Agency.

National Information Society Agency. (2019a). *Request for proposal: Integrated operation and maintenance of the public data portal and pan-government data platform*. National Information Society Agency.

National Information Society Agency. (2019b). *The development process of public data policy*. National Information Society Agency.

National Information Society Agency. (2021). *2020 survey on internet usage*. National Information Society Agency.

National Information Society Agency. (2023a). *Announcement: 2023 support project for building AI training data*. National Information Society Agency.

National Information Society Agency. (2023b). *Guidelines for quality management and construction of artificial intelligence training data v3.0*. National Information Society Agency.

National Information Society Agency. (2024a). *Request for proposal: Operation of public data portal and services (2025–2026)*. National Information Society Agency.

National Information Society Agency. (2024b). *Request for proposal: Establishment of the next-generation public data portal (Phase 1)*. National Information Society Agency.

National Information Society Agency. (2024c). *2024 big data platform-based analytics service support project: Casebook of achievements*. National Information Society Agency.

National Information Society Agency. (2024d). *Announcement: Recruitment of demand companies for 2024 big data platform-based analytics service support project*. National Information Society Agency.

National Information Society Agency. (2025a). *Guidelines for quality verification criteria of AI training data*. National Information Society Agency.

National Information Society Agency. (2025b). *Guidelines for quality management and*

verification indicators of AI training data. National Information Society Agency.

National Information Society Agency. (n.d.). *AI Hub (aihub.or.kr)*. https://www.aihub.or.kr/

National Information Society Agency. (n.d.). *Big Data Map*. http://www.bigdata-map.kr

National Information Society Agency. (n.d.). *Big Data Platform (bigdata.go.kr)*. https://www.bigdata.go.kr/

National Information Society Agency. (n.d.). *DataONAir*. https://www.dataonair.or.kr

National Information Society Agency. (n.d.). *Public Data Portal (data.go.kr)*. https://www.data.go.kr/

National Water Resources Management Committee. (n.d.). *Water resources management system (WRMS)*. https://rawris-am.ekr.or.kr/wrms/

O'Reilly, T. (2012). Government as a platform. *Innovations: Technology, Governance, Globalization, 6*(1), 13–40.

OECD. (2019). *Digital government review of Argentina: Accelerating the digitalisation of the public sector*. OECD Digital Government Studies. OECD Publishing. https://doi.org/10.1787/3c74b8c0-en

OECD. (2020). OECD digital economy outlook 2020. *OECD Publishing*. https://doi.org/10.1787/bb167041-en

OpenAI. (2023). *Introducing ChatGPT* [Blog post]. OpenAI. https://openai.com/index/chatgpt/ OpenAI

OpenAI Research. (2024). Advanced labeling strategies for AI model training. arXiv:2402.11032.

OpsNow. (n.d.). *OpsNow official website*. https://www.opsnow.com/en

Park, D. W., & Ryu, H. S. (2018). Regulatory issues and implications in the transition of the healthcare paradigm: Focusing on precision medicine and regenerative medicine. *Biomedical Law and Science, 20*, 37–62.

Park, J. S., & Park, S. A. (2003). A study on the appropriateness of informatization budget size and efficient resource allocation. *Institute of Law and Public Administration, University of Seoul*.

Park, S. (2023). How ChatGPT learns and how to use it. *SoftlyAI*. https://blog.softly.ai/how-chatgpt-learns-world/

Personal Information Protection Commission. (n.d.). *Pseudonymized data support platform*. https://www.dataprivacy.go.kr

Presidential Committee on the Fourth Industrial Revolution. (2018). *Data industry activation strategy*. Presidential Committee on the Fourth Industrial Revolution.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language*

understanding by generative pre-training.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research, 21*(140), 1–67.

Red Hat. (2022). *What are application programming interfaces?* https://www.redhat.com/en/topics/api/what-are-application-programming-interfaces

Rosset, C. (2020). Turing-NLG: A 17-billion-parameter language model by Microsoft. *Microsoft Research Blog*. https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/

Rose, T. (2016). *The end of average: How to succeed in a world that values sameness*. Penguin UK.

Russell, S., & Norvig, P. (2020). *Artificial intelligence: A modern approach* (4th ed.). Pearson.

Schwab, K. (2016). *The fourth industrial revolution* (K. J. Song, Trans.). Megastudy Books.

Seo, J. C. (2024). AGI required by the times, the journey of companies has begun. *HelloT*. https://www.hellot.net/news/article.html?no=89465

Shar, N. (2024). CUDA accelerated: NVIDIA launches array of new CUDA libraries to expand accelerated computing and deliver order-of-magnitude speedup to science and industrial applications. *NVIDIA Blog*. https://blogs.nvidia.com/blog/cuda-accelerated-computing-energy-efficiency/

Shirkey, Z., & Savic, I. (2009). *Trust in the balance: Asymmetric information, commitment problems and balancing behavior*.

Simon, H. A. (2013). *Administrative behavior*. Simon and Schuster.

Sitesbay. (n.d.). *Private cloud computing in cloud computing*. https://www.sitesbay.com/cloud-computing/cloud-computing-private-cloud-computing#google_vignette

Statista. (2023). *ChatGPT sprints to one million users: Time it took for selected online services to reach one million users* [Infographic]. https://www.statista.com

Statistics Korea, Statistical Education Center. (2019). *Securing data linkage system in the era of big data platforms*. Statistics Korea.

Susskind, D. (2020). *A world without work: Technology, automation and how we should respond*. Penguin UK.

Superb AI. (2021). *Crowdsourcing management manual for large-scale AI training data construction*. Superb AI.

Tang, A. (2019). A strong democracy is a digital democracy. *The New York Times*. https://www.nytimes.com/2019/10/15/opinion/taiwan-digital-democracy.html

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., & Lample, G. (2023). LLaMA: Open and efficient foundation language models. arXiv:2302.13971. https://arxiv.org/abs/2302.13971

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). XLNet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems, 32*.

Yoon, C. H. (2024). The evolution of AI technology and changes in the industry: AI all around #4. *SK hynix Newsroom*. https://news.skhynix.co.kr/post/all-around-ai-4

Zaki, B. K., Ko, Y., Eom, S. J., Park, K. C., Park, J. E., Cho, B., Lee, J., & Lesnichaya, Y. (2021). *Enabling data-driven innovation: Learning from Korea's data policies and practices for harnessing AI*. WBG Korea Office Innovation and Technology Notes; No. 11. Washington, D.C.: World Bank Group. http://documents.worldbank.org/curated/en/099516201122472656

Index

A

Abstract viii
Act on Promotion of Information and Communications Network Utilization and Information Protection 49
  See also Three Data Laws
Act on the Promotion of the Provision and Use of Public Data 16, 41
Administrative functions
  automation of 77, 89
  efficiency in vii, 77, 88, 96
  predictive capabilities 89, 100
Algorithm
  bias in 80, 84
  deep learning 24, 85
  fairness 84, 85, 89
  machine learning 80, 84, 98–100
  neural networks 84, 85
  transparency 84
Application Programming Interface (API) 16, 19, 21, 22, 24, 28, 29, 31, 33, 35, 36, 39, 43, 45, 56, 88, 128, 133, 136, 146
Artificial Intelligence (AI)
  ethics 72
  generative model viii, 88, 93, 107
  governance 36, 89
  history of 78
  implementation challenges viii, ix, 59
  infrastructure viii, ix, 35–37, 45
  literacy 57
  reliability 81, 89, 105
  safety 36, 45
Artificial Intelligence Hub (AI Hub)
  dataset construction 63, 64, 66, 70, 73
  sectoral domains 30

utilization cases viii, 15

B

Bias 58, 59, 70, 71, 83, 108, 109, 127
algorithmic 84
data collection 84
Big data platform
data marketplace 25
interoperability 53
sectoral domains 29

C

Categorization 133
Citizenship vii, viii, 16, 55, 56, 59, 65, 69, 77, 81, 85, 99, 100, 115, 117, 123, 125, 126, 131, 134, 136
digital vii
rights in AI era 8
Cloud computing
hybrid cloud 133
private cloud 22, 131–133
public cloud 131
Collective intelligence 64
Committee on Public Data Strategy 17, 18
Core keywords 3
Credit Information Use and Protection Act 49
See also Three Data Laws
Crowd workers
compensation 67
cost verification 66, 67
crowdsourcing 65, 66
digital skills 65, 66
quality control 51

D

Data

anonymization 79
cataloging 18, 117, 128
cleansing 78, 79, 104, 110, 111
governance viii, 7, 9, 35, 52–54, 67, 134
interoperability 21, 31, 33
labeling 34, 63, 65, 66, 72, 77, 104, 112, 127, 140
lifecycle 6, 8, 36, 53, 55, 73
openness 3, 8, 15, 18, 22, 27, 43, 50, 52, 54, 55
preprocessing 104, 138
privacy 73, 108
quality management 45, 67
security 27, 29, 37, 73
silos 101
standardization 17, 41, 45, 53, 54
Data Catalog Vocabulary (DCAT) 31, 33, 140, 141
Data Dam 8, 9
Data economy
ecosystem 8, 25, 28, 53, 54, 56, 133
market growth 5, 10, 31, 47
private-sector participation 8–10, 12, 15, 31, 33, 41, 48
Data governance
operational layer 39, 54
strategic layer 52, 56
tactical layer 52
Decision-making
data-driven 63, 85, 89
real-time 64, 101
Deep learning 24, 35, 37, 85, 87, 88, 96, 105
Digital New Deal 8, 9, 49, 63, 64, 68
Digital transformation viii, ix, 18, 54

E

E-Government vii, viii, 3, 5, 6, 49, 50, 60, 99, 101, 118–121, 123, 130–134, 137
evolution to AI Government 117, 118, 120, 131–133, 137

history in Korea viii, 3, 5, 6, 49, 50
Ethics 54
Explainable AI (XAI) 85–87
interpretability 85
local explanations 41

F

Fairness 81, 84, 85, 89
Fine-tuning 105, 108, 114, 116, 127
Foundation Model ix, 103–106, 108–110, 113, 114, 116, 134
pre-training 103–105
scalability 103
Framework Act on National Informatization 9
Future challenges 52
institutional readiness viii

G

Government AI Platform (GAIP)
architecture 146
shared services 139
Government-as-a-Platform vii
Gyeongju 27

H

Human-centric AI 77

I

Information Society 49, 51, 61, 62, 100
International cooperation
European Union (EU) 86
Organization for Economic Co-operation and Development (OECD) 6, 42, 43, 52–54, 66
United Nations (UN) 6
Interoperability 10, 21, 22, 31, 33, 53, 140
semantic 21
technical 53

L

Large language models (LLMs) viii, 35, 80, 88, 91, 93, 96, 103, 105–107, 109
Legislation
  data protection 10, 53, 89
Linked Open Data (LOD) 21, 31, 33, 139

M

Machine learning viii, 35, 44, 79, 80, 82, 84, 85, 88, 96, 98–101
Machine-readable data 20, 133, 134, 139, 140
Metadata 15, 21, 22, 35, 36, 51, 53, 117, 128, 134, 135, 140, 141, 143
  management 22, 36, 51, 53
  standards 15, 22, 140
Ministry of Science and ICT (MSIT) 3, 6, 9, 11, 12, 49, 64, 67, 68, 71, 79
MyData 54

N

National Information Society Agency (NIA) 3, 9, 10, 14, 16, 18, 19, 24, 28, 32, 34, 42, 46, 49, 64, 66, 68, 69, 71, 79, 83, 127, 144
National strategy 3
  data strategy viii, 7, 9, 140
Natural Language Processing (NLP) 79, 96, 105–107, 111

O

Open data viii, 7, 8, 16, 20, 39, 41, 43, 44, 47
  economic impact 46, 64
  high-value datasets 34, 39
Open Government Data Portal 15, 16, 19–22, 24, 39, 47, 53
  API integration 22, 39
  usage statistics 18

P

Personal Information Protection Act 29, 72, 109
  See also Three Data Laws
Personal Information Protection Commission (PIPC) 109

Platforms vii, viii, 6–8, 11, 15, 16, 18, 20, 22, 24–26, 28–33, 35, 43–45, 47–56, 65–67, 73, 78, 87, 88, 95, 96, 99–101, 110, 117, 118, 123, 125–128, 130, 131, 133, 134, 137–140, 143, 145–147
  big data viii, 6, 8, 15, 25, 26, 28–32, 43–45, 47, 53, 55, 110, 133
  government service 118, 130, 131, 137
Policy implementation 48, 49, 61–64, 66, 68, 70, 72
  correctness 63
  timeliness 61
Privacy 72, 73, 79, 108, 127
  data protection 72
Public–private partnership 48
Public data 6, 8, 9, 16–19, 21, 24, 39, 41–45, 55, 56, 68, 69, 84
  availability 43, 68
  provision process 16–18, 24, 41

Q

Quality management 34, 45, 67
  standardization 34, 45

R

Research and Development (R&D) 6, 7, 13, 26, 33, 35, 106
Resource Description Framework (RDF) 21, 31, 33, 139

S

Safe Zone 36, 37, 73
Security 4, 11, 16, 22, 26, 27, 29, 31, 37, 41, 49, 57, 73, 130, 131, 143
  cybersecurity 130
  data privacy 72, 79
Semantic Web 21
Standardization 10, 17, 22, 41, 45, 51, 53, 54
  data formats 71
  metadata guidelines 34, 44

T

Three Data Laws 49

Training data 10, 33–35, 59, 61, 69, 71, 84, 88, 98, 104, 105, 107–111, 115, 117, 143, 144
  diversity 72, 105, 108, 144
  volume 66, 105, 143
Transformer architecture 105, 106, 114
Transparency 3, 7, 15, 84–86, 89, 109
  algorithmic 89
  procedural 84

X

XAI

Table 6.6 Key hyperparameters

Hyperparameter name	Main role
Learning rate	– Determines how quickly the model attempts to reach the "correct answer" – If the learning rate is too high, the model may overshoot the optimal value; if too low, training may become slow
Batch size	– Refers to the number of data samples the model uses in a single training iteration – A larger batch size increases the amount of data processed at once, but also raises memory usage and may lead to instability during training
Number of epochs	– The number of times the model trains on the entire dataset – Determines how many times the model repeatedly learns from the data
Momentum	– An optimization technique that uses past learning information to influence the direction of current weight updates – Helps the model reach the optimal value more stably by accelerating convergence and reducing oscillations using previous gradients
Regularization parameters	– Parameters that limit model complexity to prevent overfitting, where the model fits too closely to the training data
Dropout rate	– A technique that randomly excludes certain neurons during training to prevent overfitting in neural networks
Activation function	– A function that transforms input signals to enable the model to solve complex problems beyond simple calculations
Optimizer	– An algorithm that determines how the model parameters are updated during training

Source Goodfellow et al. (2016)

Table 1.2 Overview of the Internet New Industry Promotion Plan

Category	Key details
Formation of foundation	① Preparation of laws and systems to promote the spread of new Internet industries and secure user reliability ② Promotion of R&D considering the need to secure and commercialize source technologies ③ Fostering insightful manpower by combining humanities and new Internet industry technology
Creation of market	① Creation of a sustainable ecosystem by expanding IoT services, strengthening the data and ICT resource utilization system, and supporting commercialization efforts by companies ② Promotion of leading projects to activate local industries, enhance public safety and convenience, and strengthen IT-related competitiveness
Strengthening of competitiveness	① Development of corporate support infrastructure such as open data analysis utilization center and global future Internet demonstration environment ② 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

Source Compiled by the author based on the Internet New Industry Promotion Plan

Table 2.3 Structure and data products of the 21 sectoral Big Data Platforms

Domain		Lead and operating institution	No. of participating institutions	Key data products and features
1	Finance	BC Card	11 institutions	Provides a one-stop big data service for everyday finance using both financial (card spending, insurance, securities) and non-financial (telecom, distribution, media, etc.) data
2	Environment	Korea Water Resources Corporation	12 institutions	Integrates water, climate, fine dust, ecological, and chemical data to offer services such as air quality analysis, park recommendations, and media-based environmental issue monitoring
3	Culture	Korea Culture Information Service Agency	15 institutions	Provides comprehensive cultural and leisure information services and policy dashboards by integrating data on culture, lodging, food, books, trends, and commercial areas
4	Transportation	Korea Transport Institute	12 institutions	Enhances transportation services by integrating traffic volume, public transport, navigation, real estate, and population movement data
5	Health care	National Cancer Center	10 institutions	Supports cancer diagnosis, treatment decision-making, and anticancer drug R&D by integrating clinical data for the 10 most common cancers
6	Retail and consumption	Maeil Broadcasting Network (MBN)	13 institutions	Offers lifestyle-based dining preferences and regional online shopping insights by combining retail, payment, delivery, telecom, real estate, and social media data
7	Telecommunications	KT	14 institutions	Integrates data on floating population, commercial areas, card spending, tourism, transportation card usage, and social media to provide services such as commercial district analysis and population movement insights. Offers a one-stop big data service including data supply, processing, and consulting

Domain		Lead and operating institution	No. of participating institutions	Key data products and features
8	SMEs	Duzon Bizon	12 institutions	Integrates accounting, real estate, insurance, employment, welfare, and SNS data to provide business management analysis and job demand forecasting services
9	Regional economy	Gyeonggi Provincial Government	10 institutions	Combines data on local currency usage, enterprise data, jobs, credit ratings, and demographics to deliver regional consumption pattern analysis and customized job-matching services
10	Forestry	Korea Forestry Promotion Institute	12 institutions	Merges data on forestry, hiking trails, bicycles, public transit, mountain weather, and forest disasters to provide trekking and forest product information services
11	Fire safety	National Fire Agency	9 institutions	Collects and analyzes data on firefighting, disaster safety, and the firefighting industry to provide high-quality safety information and build a data transaction foundation
12	Smart policing	Korea National Police University	10 institutions	Integrates public–private security data for risk prediction and analysis to support national public safety
13	Marine and fisheries	Korea Maritime Institute	11 institutions	Establishes a comprehensive data marketplace for marine, shipping, fisheries, and ports by integrating, processing, and distributing ocean-related data
14	Agri-food	Korea Agro-Fisheries and Food Trade Corporation	10 institutions	Uses agricultural growth stats, pork production stats, and distribution patterns to build an integrated data hub for activating agri-food data circulation and transactions
15	Lifelog	Wonju Severance Christian Hospital	12 institutions	Builds a lifelog big data ecosystem to collect, analyze, and utilize daily life records across the entire human lifecycle (from birth to death)

Domain		Lead and operating institution	No. of participating institutions	Key data products and features
16	Digital industry innovation	Korea Testing Laboratory	10 institutions	Utilizes domestic/overseas corporate data, scientific/technological and patent information to support crisis response, policymaking, and industrial innovation via data openness
17	Infectious diseases	Korea Institute of Science and Technology	10 institutions	Preserves and analyzes COVID-19 data while establishing a responsive ecosystem for future infectious diseases
18	Spatial information	Korea Land and Geospatial Informatix Corporation	10 institutions	Fosters a spatial data ecosystem and expands business opportunities by building a public–private spatial data trading platform
19	Real estate	Korea Real Estate Board	10 institutions	Fosters proptech industries by integrating housing/commercial property prices, blueprints, and transportation-environment-spatial data, supporting both public and private sectors
20	Smart farm	Naver Cloud	10 institutions	Promotes digital agriculture based on carbon–neutral data processing, livestock disease diagnostics, and distribution-enhancing smart farm data
21	Coastal zone	Korea Institute of Ocean Science and Technology	10 institutions	Establishes a data ecosystem for national disaster and safety management by using coastal weather, terrain, climate, and marine debris data to support public–private integration and new industry development

Source National Information Society Agency (2024c)

Table 6.4 Comparison of major labeling methods for unstructured data

Labeling method	Definition	Use cases
Key point	A labeling method that designates specific points of an object within an image using coordinates	Human pose estimation, facial landmark detection, motion analysis
Optical character recognition (OCR)	Labels the location and content of text within an image to create data for text recognition	Document digitization, road sign recognition, license plate recognition
Semantic segmentation	Assigns every pixel in an image to a specific class	Autonomous driving (road, vehicle detection), medical imaging (tumor detection)
3D cuboid	Labels position, size, and rotation by defining an object as a 3D cuboid (rectangular box)	Autonomous driving (vehicles, pedestrians), robotic vision, logistics automation
Bounding box	The basic labeling method that defines the location and size of an object using a 2D bounding box	Object detection (vehicles, people), object classification

Source OpenAI Research (2024b)

Table 7.2 Government AI-based services vs. Government AI Platform

Category	Government AI-based services	Government AI Platform
Definition	Application systems built by the government using AI technology to deliver specific public services	A set of technical foundations and tools that support the development, operation, management, and scaling of AI systems
Purpose	Performs specific functions (e.g., civil service processing, policy recommendation) for end users—citizens and public institutions	Provides an environment and infrastructure for developing, deploying, training, and managing public AI systems efficiently
Users	– Citizens (users of government services) – Public institution staff (administrative personnel)	– Developers of AI systems in the public sector – Data scientists, ICT companies – Government AI research teams and partner organizations

Source Ministry of the Interior and Safety (2022)

Table 6.5 Key evaluation metrics for labeled data

Metric	Description
Accuracy	– Proportion of correctly labeled data out of the total dataset
Precision	– Proportion of predicted "positive" labels that are actually correct
Recall	– Proportion of actual "positive" cases that the model correctly labeled as positive
F1 score	– Harmonic mean of precision and recall
Labeling consistency	– Measures how consistent the labeling results are when multiple annotators label the same data
Noise rate	– Measures the proportion of incorrectly labeled data to assess overall data quality

Source Géron (2022)

Table 1.4 Key contents of the Data Industry Promotion Act

Category	Key provisions
① **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. Related terms such as "data" are also defined
② **Development of basic plan**	The government devises a basic plan for the promotion of the data industry every three years, covering data production, distribution, utilization, and foundation-building for the data industry
③ **National Data Policy Committee**	A committee under the Prime Minister that deliberates and coordinates major national data policies (e.g., utilization, security, circulation, and transactions). Responsible for reviewing and coordinating essential policies and plans
④ **Protection of data assets**	Protects data that have economic value ("data assets") and are created through substantial human and material investment. Unauthorized acquisition, use, disclosure, provision, or removal of technical protection from data assets without legal authority is prohibited
⑤ **Data valuation support and quality control**	Prepares valuation techniques, systems, certification targets, and standards for data. Also designates valuation and quality certification agencies responsible for related tasks
⑥ **Data trader**	A person with expert knowledge in data trading can register as a data trader with the Ministry of Science and ICT (MSIT), which provides training and information necessary for performing data trading tasks
⑦ **Data industry promotion and business report**	Data trading businesses, data analysis providers, etc., must report to MSIT. MSIT and related central agencies may provide financial, technical, and administrative support as needed
⑧ **Data industry support (startups, SMEs)**	When implementing data-related support policies (e.g., commercialization, data capability enhancement), small and medium-sized enterprises (SMEs) are prioritized. Partial support for necessary costs such as data processing and trading is also provided
⑨ **Fostering of professional workforce**	MSIT and the Ministry of Personnel Management (MOPAS) develop policies to foster data professionals. MSIT designates and supports professional training institutions

Category	Key provisions
⑩ **Establishment of data distribution system**	Establishes a system for supporting data distribution and transactions, such as data voucher programs and data distribution platforms

Source Ministry of Science and ICT (2021)

Table 7.5 Components of a Whole-of-Government Service Catalog (example)

Category	Details
Service list	A list of all services developed and operated by all ministries and agencies of the government
Metadata	Detailed information for each service, such as service name, service ID, service description, and service provider
Service classification system	A structured standard for systematically classifying and organizing government services based on governmental functions and service types
Service API	A function that allows one service to be accessed and used by other external services

Source Korea Ministry of Government Legislation (2025)

Table 2.1 Key provisions of the public data act

Category		Description
General provisions	Purpose and definitions	– Purpose: To guarantee the public's right to access public data – Provision: Data must be delivered or made accessible in a machine-readable format – Machine-readable format: A format in which public data can be processed—edited, converted, or extracted—by software
	Basic principles	– Public institutions shall take necessary measures to universally expand the right to use public data and ensure its commercial utilization
Policy formulation	Public Data Strategy Committee	– A committee under the Prime Minister's Office to deliberate on major policies – Deliberates and decides on matters such as formulation and revision of the master plan, and the scope of data to be provided
	Master plan implementation and evaluation	– Formulation of a master plan to promote public data provision and utilization (every 3 years by the Ministry of the Interior and Safety) – Annual implementation plans by central and local governments – Evaluation of public institutions' performance in providing public data (by the Ministry of the Interior and Safety)
	Data Officers and Support Centers	– Designation of responsible officers in each institution – Establishment of support centers to assist in public data operations
Infrastructure development	Public data and related information subject to provision	– Target data: Relevant data held by institutions (excluding classified or copyrighted content) – Deliberation (Strategy Committee): Review and decision on the list of data to be provided – Data registration: Via the Public Data Portal (by public institutions) – Data preparation: Public institutions must ensure that the data they provide through various methods are in machine-readable format – Infrastructure support (by the Ministry of the Interior and Safety): Portal operation, quality control, standardization, and data curation support

Category		Description
Data provision	Provision and dispute mediation	– Methods: Direct provision or via the Public Data Portal – In case of refusal or suspension of data provision, a dispute mediation committee may be convened
Supplementary provisions	Immunity clause	– Public officials are exempt from civil or criminal liability for damages caused by the suspension of data provision

Source Authors' own work

Table 4.2 Four essential quality verification indicators for AI training data

Indicator	Description
Diversity	Ensures the dataset includes various situations, classes, and contexts to prevent bias
Syntactic accuracy	Verifies whether the data is grammatically or structurally correct according to format and rules
Semantic accuracy	Checks whether the content and meaning of the data are consistent and logically sound
Validity	Confirms whether the data are relevant, usable, and suitable for training the target AI model

Source Ministry of Science and ICT & National Information Society Agency (NIA) (2023b)

Table 6.3 Representative methods for data deduplication

Deduplication method	Description
Similarity-based duplicate detection	Identifies documents containing similar content—i.e., documents with repeated or redundant information—and retains only one while removing the rest
Hash value comparison	Generates a unique hash value for each document and deletes duplicates that share the same hash
Dynamic filtering	Removes unnecessary information in real time when data are continuously changing or being updated
Source tracking of duplicate documents	Traces the origin of duplicated documents; if the same content appears across multiple sites, the original is prioritized and the copies are removed
Utilization of NLP techniques	Detects documents that convey the same meaning using different expressions and removes semantically redundant content

Source OpenAI Research (2024b)

Table 8.2 Detailed indicators by data quality measurement category (examples)

Indicator	Key description
Diversity	Class distribution, instance distribution, sentence length, vocabulary size, etc.
Syntactic accuracy	Consistency between attribute values of labeled data and the originally defined data format and input value ranges
Semantic accuracy	Accuracy, precision, recall
Validity	Performance improvement after data training, measured by accuracy, precision, recall, etc.

Source National Information Society Agency (NIA) (2025b)

Table 5.7 Key functions of an AI assistant (examples)

Category	Detailed description
Chatting with AI	– Multi-chatbot: Converse with various LLMs (GPT-3.5, GPT-4, Bard, Claude) in a single interface – Prompt* Library: Quickly access and use stored prompts – Real-time Info: Retrieve current internet data during conversation (e.g., "What's the weather now?") – Voice Support: Chat using a microphone without keyboard input
Chat and summarization	– ChatPDF: Upload PDFs and chat with the document for better understanding – Image + Chat: Upload images and ask questions using GPT-4 V – YouTube summary: Get summaries of videos without watching the full content
Search	– Search agent: Uses multiple search terms to find, review, and respond – Enhanced search: Shows ChatGPT responses alongside engines like Google or Bing for easy access
Writing	– Compose tool: Customize and generate essays/reports quickly with adjustable length, style, and tone – Writing agent: Automatically generate outlines with expanded content and references based on a given topic – Email reply: Draft responses in Gmail based on email content
Translation	– PDF translation: Translate PDFs and compare original (left) vs. translation (right) – Parallel translation: Translate pages without hiding original text for easy comparison – Text translation: Instantly translate selected text on web pages
Creative arts	– Text-to-image: Convert written text into visual images
AI notes	– AI knowledge base: Save webpages, chat logs, images, and PDFs; chat with notes to access stored information – Personalization: More personalized and accurate responses as content grows

* A prompt refers to the input given to generative AI, such as questions or instructions typed into ChatGPT. Well-crafted prompts can lead to higher-quality outputs, which is why building a library of effective prompts and their results can help users receive faster and more accurate responses to similar queries in the future. The practice of

creating effective prompts is called prompt engineering, while the act of cleverly manipulating prompts to bypass AI restrictions (e.g., on unethical content) or to reveal the AI's identity is referred to as prompt hacking

Source https://monica.im/

Table 5.8 Examples of commercialized AI-based professional services

Sector	Details
Healthcare	• IBM's Watson Assistant AI Healthcare Chatbot – A conversational AI solution built on deep learning, machine learning, and natural language processing (NLP) models. It understands user queries, retrieves optimal responses, and completes transactions – Supports administrative efficiency, reduces medication errors, assists with virtual nursing tasks, and helps manage post-operative pain
Finance	• BloombergGPT: A large language model specialized in finance – built from the ground up with 50 billion parameters tailored for financial tasks – First, it accesses and analyzes financial data on the Bloomberg platform, enabling tasks such as data retrieval, analysis, report generation, and insight discovery – Second, it provides headline suggestions for news articles—helping journalists craft engaging and informative newsletters by generating relevant and compelling titles from input paragraphs – With continued training and optimization on financial data, BloombergGPT is expected to expand its capabilities across a broader range of financial operations and ultimately support more accurate and efficient decision-making in the financial industry

Source Bloomberg (March 3, 2023)

Table 8.1 Quality measurement indicators for data used in training government AI models

Indicator	Key description
Diversity	Verifies whether the quantity per category (or object) is balanced and appropriately proportioned
Syntactic accuracy	Verifies whether the planned data structure, format, and input value ranges are followed
Semantic accuracy	Verifies how accurately and precisely the labeling has been performed
Validity	Verifies the performance improvement of the AI algorithm after training on the constructed data

Source National Information Society Agency (NIA) (2025a)

Table 7.1 Rule-based system vs. AI-based system

Category	Rule-based system	AI-based system
Operating principle	Operates based on explicitly defined "if–then" rules	Learns from data and generates or optimizes rules autonomously
Design method	All rules are written and managed by humans	The model learns from data and automatically extracts patterns and rules
Inference method	Decisions are made based on logical rules	Uses probabilistic inference and pattern recognition

Source Russell and Norvig (2020)

Table 1.3 Overview of the Data Industry Activation Strategy

Category	Key details
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 workforce

Source Compiled by the author based on the Data Industry Activation Strategy

Table 2.2 Key implementation framework of public data policy

Category		Description
Policy	Public Data Strategy Committee	Deliberates, coordinates, and evaluates key government policies and plans related to public data
	Ministry of the Interior and Safety	Oversees overall management, sharing, disclosure, and utilization of public data
Support	Public Data Utilization Support Center	Supports efficient provision and active use of public data
	Public Data Provision Dispute Mediation Committee	Mediates disputes related to refusal or suspension of public data provision by public institutions
Execution	Public Data Provision Officer	Oversees tasks related to data provision and promotes data use within each institution

Source Compiled by the author with reference to the Public Data Act

Table 1.1 Overview of the big data master plan and big data service activation strategy

Category	Key details
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 of manpower	Industry-academia-research joint research and development

Source Compiled by the author based on respective plan documents

Table 3.1 Performance of national core data openings (Phases 1–4)

Phase	Years	Released data examples
1st	2015–2016	33 datasets including building information, actual real estate transaction data, and commercial district data
2nd	2017–2019	63 datasets including spatial convergence data, medical imaging data, and multilingual voice data
3rd	2020–2022	73 datasets including autonomous driving integrated control, 119 emergency dispatch information, and industrial accident decision documents
4th	2023–2025 (in progress)	30 datasets including passport authenticity verification service and franchise disclosure documents (ongoing)

Note As of the end of December 2023
Source Ministry of the Interior and Safety (2024)

Table 5.2 AI learning methods

Learning type	Method description	Example
Supervised learning	AI is trained using labeled data where the correct answers are provided	AI is trained with many labeled photos of dogs and cats, then learns to distinguish them
Unsupervised learning	AI is trained without labeled data, finding patterns, similarities, or groupings in the data itself	AI learns from many unlabeled images and groups similar ones (e.g., dogs vs. cats)
Reinforcement learning	An agent interacts with a defined environment, receiving rewards or penalties to optimize decisions	Used in AlphaGo: the system learns Go rules and finds optimal moves through trial and error

Source Author

Table 6.2 Representative open-source large language models (LLMs)

Model name	Parameters	Developer	Details
LLaMA 2	7B–70B	Meta	– Free for commercial use – Performs general NLP tasks
Mistral 7B	7B	Mistral AI	– Efficient parameter structure – Designed for large-scale AI systems
GPT-NeoX 20B	20B	EleutherAI	– Open-source large-scale language model – Supports generative AI and translation
BLOOM	176B	BigScience	– Multilingual large model – Suitable for translation and text generation
OPT	175B	Meta	– Open-source model similar to GPT-3 – Designed for research and applied use
Falcon LLM	7B, 40B	Technology Innovation Institute (TII)	– High efficiency relative to performance – Supports generative AI and document summarization
Dolly	–	Databricks	– Free for commercial use – Built for conversational systems and content generation
RedPajama	2B, 7B, 20B	Together	– Open-source and scalable – Intended for research and experimentation

Source Meta AI et al. (2023–2024)

Table 5.3 Examples of explainable AI applications by sector

Sector	Companies and use cases
Healthcare	(IBM Watson) Analyzes patient medical data and presents the reasoning behind diagnostic and treatment recommendations to help healthcare professionals understand the AI's decisions Example: Explains how a specific genetic mutation is related to cancer development and which academic papers and research findings support the proposed treatment
Finance	(KakaoBank) Applied to detect abnormal transactions, evaluate credit scores and loan eligibility, recommend personalized financial products, and support AI-based investment decisions Example: While AI provides the credit score result, XAI explains that income contributed A%, existing loans B%, and delinquency history C%—making the score transparent to the customer
Manufacturing	(Siemens) Uses sensor data in factories to predict maintenance schedules via XAI Example: When predicting a machine part replacement, XAI explains that vibration levels increased by 20% and an anomaly was detected under specific temperature conditions—supporting the replacement decision

Source Compiled by the author with reference to official service websites and Electronic Times (2024)

Table 5.5 Differences between artificial general intelligence (AGI) and narrow artificial intelligence (Narrow AI)

Category	Artificial general intelligence (AGI)	Narrow artificial intelligence (Narrow AI)
Definition	AI that demonstrates intelligence comparable to that of humans	AI that is specialized in limited tasks or specific domains
Capability	Mimics the full range of human intelligence functions such as learning, reasoning, and problem-solving	Performs functions optimized for specific tasks
Scope of application	Can be flexibly applied across various fields	Applicable only to a narrow range of tasks or problems
Autonomy	Capable of autonomous decision-making at a human level	Makes decisions based only on pre-programmed rules or specific data
Learning ability	Learns generalized knowledge from limited data and applies it to new situations	Optimized through learning from large amounts of data or task-specific training
Creativity	Has the potential to solve new problems and perform creative tasks	Focused on solving given problems; limited capacity for creative work
Societal and ethical impact	Potential to have wide-ranging impacts across all aspects of human society	Significant impact in specific fields, but less broad than AGI

Source Samsung SDS (2024)

Table 5.4 Examples of global cloud services for AI

Major global cloud services	AI services in use
NVIDIA DGX Cloud	– NVIDIA Riva: Real-time speech recognition (STT), text-to-speech (TTS), translation
AWS (Amazon Web Services)	– Claude by Anthropic: Conversational AI chatbot – SageMaker: Cloud-based machine learning platform for training and deploying AI models
Google Cloud	– Cohere for Enterprise: Search, Q&A, document summarization – Vertex AI: Comprehensive AI development platform
Meta (RSC)	– Research SuperCluster: Used for training Large Language Model Meta AI (LLaMA
Microsoft Azure	– ChatGPT via Azure OpenAI API: Generative conversational AI – Azure AI Services—Cognitive services, ML tools, and OpenAI API support
IBM Watson	– Watson Health, Watson Assistant—Healthcare AI, chatbots, enterprise AI applications

Source Compiled by the author

Table 7.3 Datasets used to train the base model of ChatGPT

Dataset name	Tokens	Ratio	Size (GB)
CommonCrawl (website data)	410B	1:1.9	570
WebText2 (social media data)	19B	1:2.6	50
Books1 and Books2 (E-books, papers, academic data)	12B	1:1.75	21
	55B	1:1.84	101
Wikipedia	3B	1:3.8	11.4
Total	499B		753.4

Source Brown et al. (2020)

Table 7.4 Data requirements for creating Government AI fine-tuning models

Category	Details
Unbiased	Balanced data that include diverse sources and topics, free from bias toward specific groups or subjects
High quality	Thoroughly cleaned and verified data with no labeling errors, noise, or missing values
Privacy protection	Data in which personal information such as names, addresses, resident registration numbers, and facial images has been anonymized

Source National Information Society Agency (NIA) (2025b)

Table 5.6 Stages of general AI development and examples of major services

Performance (row) × Generality (column)	**Performance (narrow)** **Clearly defined and limited in scope**	**Generality (general)** **Various non-physical tasks including metacognitive activities such as learning new skills**
Level 0—NO AI	– No narrow AI – Basic software such as calculators and compilers	– No general AI
Level 1—Emerging (unskilled worker level)	– Emergence of Narrow AI – GOFAI (good old-fashioned AI) (Boden, 2014) – Simple rule-based systems, such as SHRDLU	– Emergence of General AI – ChatGPT (OpenAI, 2023) – Bard (Google)
Level 2—competent (mid-level skilled worker or higher)	– Competent Narrow AI – Toxicity detectors (Das et al., 2022) – Smart speakers like Siri (Apple), Alexa (Amazon), or Google Assistant – Vision-and-language systems such as PaLI (VQA systems) – IBM Watson – State-of-the-art large language models (LLMs) capable of limited tasks, e.g., short essay writing or simple coding	– Competent AGI – Not yet achieved
Level 3—Expert (Top 10% skilled worker level)	– Expert Narrow AI – Spelling and grammar checkers like Grammarly – Generative image models such as Imagen and DALL·E 2	– Expert AGI – Not yet achieved
Level 4—Virtuoso (Top 1% skilled worker level)	– Virtuoso Narrow AI – Deep Blue – AlphaGo	– Virtuoso AGI – Not yet achieved
Level 5—superhuman (beyond the best human capabilities)	– Superhuman Narrow AI – AlphaFold, AlphaZero, Stockfish	– Artificial Superintelligence (ASI) – Not yet achieved

Note This table represents a horizontal, matrix-based approach to categorizing the stages of general AI development based on performance (rows) and generality (columns). Current systems are approximately mapped as illustrative examples
Source Morris et al. (2024)

Table 4.1 Verification system for cost processing of crowd workers

Step	Detailed description
Step 1	Mandatory use of the government-provided standard contract for all crowd workers
Step 2	Monthly full inspection of deliverables from high earners (those receiving over two million KRW)
Step 3	Random verification of individual work results through platform log records
Step 4	Operation of a reporting website for suspected fraudulent cost claims related to crowd workers
Step 5	Mandatory submission and full verification of individual payment records based on work volume

Source Superb AI (2021)

Table 3.2 Public data services by user type

User type	Service type	Main features
Experts	Catalog service	– Detailed information on datasets provided by data platforms – API-based data provision
	Update notification	– Notifications via SMS/email for changes to subscribed datasets or API services
	Advanced search service	– Advanced search options such as keyword association, inclusion/exclusion filters – Integrated search across public data portals and information disclosure sources
General citizens/public officials	Data recommendation	– Suggests commonly searched datasets – Auto-complete function
	AI-based search service	– Recommends data and analytical methods tailored for beginners based on analysis goals and domains

Source Author's compilation

Table 1.5 National Data Policy's implementation system

Institution	Functions
National Data Policy Committee	Oversees both public and private-sector data policies (chaired by the Prime Minister) Deliberates on the formulation and coordination of basic plans and strategies to promote the data industry Reviews policy and regulatory improvements related to data production, transaction, and utilization
Ministry of Science and ICT (MSIT)	Performs overall duties related to the planning, coordination, and management of National Data Policy
Data Dispute Mediation Committee	Mediates disputes related to the production, transaction, and utilization of data
Implementing institutions	Integrated Support Center [NIA] Specialized support agencies Data exchanges Specialized agencies for pseudonymized data integration Data Safe Zones Data valuation institutions

Source Compiled by the author based on the Framework Act on Data

Table 6.1 Representative foundation AI models

Developer	Model name	Model characteristics
OpenAI	GPT	A general-purpose language model based on the Transformer architecture, pre-trained on large-scale data. It performs a wide range of natural language processing (NLP) tasks with high accuracy and excels at creative text generation and contextual understanding
Google	BERT	A pre-trained language model using bidirectional (contextual) language modeling on the Transformer architecture. It demonstrates strong performance across various NLP tasks by deeply understanding context
Meta	LLaMA	A family of models based on the Transformer architecture designed to deliver efficient learning and performance even with fewer resources. Released in multiple sizes (7B, 13B, 30B, 65B) as open source for broad research and development use
Google	T5	A pre-trained language model optimized for a wide range of NLP tasks such as translation, summarization, and question answering, showing excellent performance
Google and Carnegie Mellon University	XLNet	Designed to overcome limitations of BERT, offering more sophisticated contextual understanding and superior performance in a variety of NLP tasks
Microsoft (MS)	Turing-NLG	Based on the Transformer architecture, this model delivers strong performance in natural text generation, translation, and summarization. It particularly excels in conversational responses and creative content generation
BigScience Project	BLOOM	A large-scale multilingual language model with 175 billion parameters. It supports various languages and tasks, and is released as open source to promote accessible research and collaborative advancement in NLP

Source Brown et al. (2020), Devlin et al. (2018), Meta AI Research (2023), Raffel et al. (2020), Yang et al. (2019), Rosset (2020), and BigScience Workshop (2022)

Table 5.1 Data processing procedure

Data processing procedure	Description
Data collection and storage (raw data)	Activities to secure "raw data" necessary for AI machine learning, either by generating it directly from the real world or by collecting it from existing organizations or systems that already possess relevant data
Data cleaning (source data)	– The acquired raw data are processed to fit the required format and size for machine learning, with duplicate entries removed and any personal information anonymized. Through this process, source data are secured – The source data at this stage refer to data that have not yet been labeled. It is refined raw data that have undergone the necessary cleaning steps in preparation for the labeling process
Processing and verification (labeled data)	– A labeling process is performed by attaching labels to the source data that align with the intended function or purpose, making the data usable for machine learning – The resulting labeled data refer to a set of ground-truth files assigned to the source data, including annotations that contain data attributes such as format and resolution, along with descriptions and comments – A verification process is carried out based on the type of data—such as raw data, source data, or test data—to ensure its quality and suitability for training

Source Author's compilation based on Ministry of Science and ICT & National Information Society Agency (NIA) (2023b)