

Translation



The following draft Chinese government policy is designed to increase the quantity and quality of datasets available to train AI models for various industries. The draft addresses data as a growing bottleneck for AI development in China, particularly as regards datasets that can be used to train and test AI systems that interact with the physical world, such as AI-enabled robots. The authors invited public feedback on the draft from April 15 to 22, 2026. China subsequently released the final version of this policy on June 8, 2026; a CSET English translation of the final version is underway.

Title

Implementation Plan for the Initiative to Promote the Construction of High-Quality Industrial Datasets (Draft for Feedback)
关于推进行业高质量数据集建设行动的实施方案(征求意见稿)

Author

Integration Department of the Chinese National Data Administration (国家数据局综合司)

Source

National Data Administration website, April 15, 2026.

The Chinese source text is available online at:

https://www.nda.gov.cn/sjj/hdjl/yjq/yjqform/list/index_pc.html?code=ff808081-9b5b91e0-019d-9170882a-3b0f

An archived version of the Chinese source text is available online at: <https://perma.cc/V4DA-DY39>

Translation Date

July 15, 2026

Translator

Etcetera Language Group, Inc.

Editor

Ben Murphy, CSET Translation Manager

Implementation Plan for the Initiative to Promote the Construction of High-Quality Industrial Datasets (Draft for Feedback)

High-quality industrial datasets are collections of industry data that have undergone data handling, including collection, processing, and other operations; can be used directly to develop and train artificial intelligence (AI) models; and can effectively improve the effectiveness of applications such as models, intelligent agents, and smart terminal devices. They include industrial general knowledge datasets and industrial specialized knowledge datasets. High-quality industrial datasets are foundational and key resources for promoting “AI+” to empower all industries and achieve industrial

implementation. This plan is formulated in order to implement the *Outline of the 15th Five-Year Plan for National Economic and Social Development*, thoroughly implement the “AI+” initiative, promote the construction and adoption of high-quality industrial datasets in step with, and in mutual reinforcement of, “AI+,” and strengthen the role of data in empowering the innovative development of AI.

1. Overall Requirements

Guided by Xi Jinping Thought on Socialism with Chinese Characteristics for a New Era, we will thoroughly implement the spirit of the 20th Party Congress and all subsequent plenums, comprehensively implement the “AI+” initiative, proactively adapt to the paradigm shift in AI development, and follow the principles of “demand-driven, urgent needs first, application validation, and security assurance.” Focusing on key industries for national economic development and strategic emerging industries, and centering on key links such as the supply, circulation, and application of high-quality industrial datasets, we will launch six special initiatives: (1) Strengthening foundations and expanding capacity, (2) tackling the challenges of data labeling, (3) improving quality and efficiency, (4) empowering applications, (5) management services, and (6) creating value. This will form a “data flywheel” in which “scenarios lead the way for data, data drives models, models empower applications, and applications create value,” and accelerate the building of a symbiotic ecosystem in which the data factor of production (数据要素) and AI evolve in coordination.

By the end of 2028, a number of high-quality industrial datasets that cover key fields and have been validated through application will have been established; a number of typical application scenarios in which data drives the innovative development of AI will have been created; a number of innovative data enterprises and professional personnel with leading advantages will have been cultivated; and a number of standards and tools for building high-quality industrial datasets will have been developed. A virtuous cycle from data supply to value creation will have basically taken shape, the role of data in empowering the innovative development of AI will become more pronounced, the data industry and AI will be deeply integrated, and new growth points in the intelligent economy will continue to emerge.

2. Implement the Initiative to Strengthen Foundations and Expand Capacity

In response to the trend of AI’s accelerated penetration into industries and its paradigm shift from dialogue toward multimodal generation, decision-making and execution, embodied AI, physical interaction, and other areas, broaden data supply channels, enrich the types of data supplied, and accelerate the building of high-quality industrial datasets to provide sufficient “fuel” for AI development and application.

(1) Focus on industry fields to advance the building of high-quality datasets.

Focus on key fields such as scientific research, industrial manufacturing, agriculture and rural areas, smart energy, transportation, financial services, healthcare, education and teaching, e-commerce, human resources, culture and tourism, emergency management, meteorological services, green and low-carbon development, public security, urban governance, housing construction, and natural resources, as well as innovative fields such as the low-altitude economy, embodied AI, intelligent driving, smart ocean, and biomanufacturing, and accelerate the building of high-quality industrial datasets.

(2) Systematically advance the building of high-quality industrial datasets.

Take stock of industry data resources and application scenarios, and establish inventories of data resources and dataset demand. While being led by applications, continue to advance pilot projects for building high-quality industrial datasets, and accelerate the formation of a number of replicable and scalable data-driven demonstration scenarios. Strengthen the leading and driving role of production chain-leading (链主) work units (单位), promote collaborative development and resource integration across the upstream and downstream of production chains, continuously expand the supply scale of high-quality industrial datasets, and encourage lead production chain entities to open datasets to their industries and provide data services. Intensify the development and utilization of public data resources, and build a number of high-quality datasets based on public data.

(3) Advance the building of high-quality industrial datasets in line with AI development. For all stages of AI pretraining, instruction fine-tuning, reinforcement learning, evaluation, and other processes, continue to advance the building of multimodal high-quality datasets for text, code, images, audio, video, point clouds, time-series data, scientific data, and other data types. For intelligent agents and other new forms of intelligent applications, strengthen the building of datasets such as knowledge bases, knowledge graphs, and ontologies, and accelerate the building of datasets for complex task planning, long-range reasoning, human-computer interaction, decision-making and execution, and other tasks. In order to meet the development needs of embodied AI, accelerate the building of datasets related to interaction between machines and the real world (真机交互) for physical interaction, environmental perception, motion control, and other functions in key scenarios, and actively apply emulation, simulation, and synthesis technologies to expand the data supply. Actively plan the building of datasets for cutting-edge directions such as world models.

(4) Strengthen organic linkage with the construction of data infrastructure.

Encourage reliance on national data infrastructure, make full use of capabilities such as

privacy-preserving computing and trusted data spaces, carry out secure dataset storage, trusted circulation, and efficient application, and promote the transformation of datasets from dispersed holding to intensive and standardized supply. Encourage exploration of the construction of data infrastructure storage capacity centers that support large-scale, multimodal datasets.

3. Implement the Initiative to Tackle the Challenges of Data Labeling

Data labeling is the process of injecting knowledge and experience into training data, and is an indispensable key link in the building of high-quality industrial datasets. Guide data labeling in its transition from being primarily human-based to a multi-level labeling model featuring human-computer coordination and deep expert participation (专家深度参与), and promote a leap in data labeling toward professionalization, intelligentization (智能化), and systematization.

(5) Promote the transformation and upgrading of data labeling. Strengthen scientific and technological (S&T) innovation in the field of data labeling, strengthen the R&D and application of automated tools and platforms, develop intelligentized labeling services such as “model pre-labeling + manual calibration,” “manual labeling + model verification,” and “model pre-labeling + model verification,” and comprehensively improve data labeling capabilities. Develop expert data labeling services, establish an industry expert certification and registration mechanism, promote deep expert participation in the professional knowledge labeling needed for stages such as instruction fine-tuning and reinforcement learning, produce high-quality datasets for logical reasoning, domain knowledge, and other areas, and increase the knowledge density and professional value of datasets.

(6) Continue to promote pilot and trial work on data labeling. Guide the first batch of seven cities undertaking pilot and trial tasks for data labeling to continue strengthening and deepening the data labeling industry. For regions with strong innovation capacity, a sound development foundation, and strong industrial characteristics, make phased arrangements for a number of data labeling innovation pilot zones. Guide qualified regions in an orderly manner to carry out pilot zone development in light of local conditions, avoid blindly following the latest trend (一哄而上), prevent homogenized, low-level redundant construction, promote close collaboration across the upstream and downstream of the data labeling production chain, and form industrial clustering effects. Cultivate a number of leading data labeling enterprises, unicorn enterprises, and gazelle enterprises, and expand the data labeling industry.

(7) Expand the supply of data labeling talent. Support qualified colleges and universities in adding courses related to data labeling, and rely on industry-education

integration, school-enterprise collaboration, and other approaches to cultivate data labeling talent with specialized knowledge. Encourage occupational skill level recognition for data labeling, strengthen continuing education and on-the-job training, and open up pathways for talent development. Improve the tiered and categorized talent evaluation system, and build a professional labeling talent cadre that combines full-time and part-time personnel. Encourage college and university graduates and other groups to participate in labeling work, and expand employment channels through flexible employment, project collaboration, diversified job supply, and other approaches.

4. Implement the Initiative to Improve Quality and Efficiency

Promote the construction of high-quality datasets that meet quality standards such as structural integrity, content diversity, labeling accuracy, and model adaptability, and that satisfy AI readiness requirements, so as to reduce training and inference costs and effectively improve model performance.

(8) Improve the quality and efficiency of high-quality industry dataset construction. Strengthen efforts to tackle key technologies such as data cleaning, augmentation, labeling, alignment, and quality inspection, as well as R&D, and application of automated tools for the entire process, to support the efficient, high-standard construction of high-quality industrial datasets. For specific industry application scenarios, encourage the use of technologies such as intelligent data filtering and blending to build more refined and more powerful high-knowledge-density datasets and reduce training and inference costs. Give full play to the positive role of data synthesis in dataset construction, and use data generated by models, simulation systems, and other means to solve problems such as the difficulty of constructing datasets for data-scarce scenarios and the high cost of collecting data from real-world scenarios.

(9) Promote the development and implementation in practice of a standards system for high-quality datasets. Accelerate the formulation of relevant national standards for the formats, types, labeling, quality evaluation, and other aspects of high-quality datasets, and improve the standards system for high-quality datasets. Encourage all industries and localities to coordinate with national standards, and promote the formulation of standards for high-quality datasets in key industry fields. Strengthen standards application validation and standards implementation, and promote the standardized construction of high-quality datasets.

(10) Strengthen quality evaluation of high-quality datasets and mutual recognition of results. Continue to improve the “data quality validation + model application feedback” evaluation method, accelerate the building of evaluation datasets

covering multiple industries, scenarios, and modalities, and effectively evaluate the application results of high-quality datasets. Launch a joint initiative for high-quality dataset evaluation, encourage relevant work units to carry out evaluation and packaging work in accordance with standards and by adopting unified evaluation plans and tools, and achieve “one evaluation, nationwide mutual recognition.”

5. Implement the Initiative to Empower Applications

Persist in deeply integrating high-quality industry dataset construction with practical application, use models to draw out data and use data to empower models (以模引数、用数赋模), promote the building of high-quality datasets in step with, and in mutual reinforcement of, the “Data Factor of Production ×” and “AI+” initiatives, and comprehensively empower the digitalized and intelligentized (数智化) transformation of industries.

(11) Build an application closed loop for the “data flywheel.” Use model applications to guide data supply and data to drive model iteration, creating a virtuous cycle in which “scenarios—data—models” develop in coordination. Give play to the driving role of “AI+” scenarios, promote precise matching between data supply and scenarios, promote construction through application, attract the aggregation of more data resources through practical demand, and promote the effective supply and continuous optimization of high-quality industrial datasets. In conjunction with the “Data Factors of Production ×” initiative, deepen the building of high-quality industrial datasets, promote application through construction (以建促用), fully utilize dynamic interaction data generated by model applications, and drive the continuous improvement of industry model capabilities.

(12) Create industry application benchmarks and demonstration projects. Focus on creating data empowerment workshops (数据赋能工场) that integrate “dataset production, processing, circulation, and utilization” with “support for model training and application,” create a number of industry benchmarks, and accelerate the implementation of AI applications. Create a number of “data × intelligent agent” demonstration projects, establish models for the successful use of high-quality datasets to drive intelligent agents in solving actual problems, and lead and drive large-scale application.

(13) Foster a thriving ecosystem for collaborative dataset development. Strengthen overall coordination, enhance departmental linkage, and advance the building and application of high-quality datasets in an orderly manner by both industry and field. Build platforms for exchanges among enterprises, universities and colleges, scientific research institutions, industry associations, financial institutions, and other parties, and form an industrial ecosystem featuring joint development and sharing,

mutual benefit, and win-win outcomes. Regularly hold supply-demand linking activities, improve the efficiency of supply-demand matching, and promote the implementation of substantive cooperation. Organize the selection of typical cases in the construction and application of high-quality industrial datasets, and give play to their role in providing industry demonstrations and guidance. Support the holding of innovation competitions and other activities for high-quality industrial datasets, and use competitions to promote construction and application. Deepen international exchanges and cooperation, and encourage entities engaged in building high-quality industrial datasets to actively participate in the development of the global data ecosystem.

6. Implement the Management Services Initiative

Strengthen dataset management, improve data ethics and governance mechanisms, promote the implementation of systems related to data rights and interests, and make the dataset construction system more standardized and orderly.

(14) Build a full-lifecycle management system for datasets. Strengthen the development of dataset management service capabilities covering the total life cycle, including data collection, processing, labeling, quality inspection, evaluation, iteration, and auditing. Relying on data infrastructure, strengthen the application of technologies such as privacy-preserving computing and blockchain to ensure that data is manageable, controllable, and traceable. Build a national dataset management platform that is “physically distributed and logically centralized,” and enable interconnectivity and interoperability of information such as dataset catalogs and supply and demand. Support localities and industries in setting up dedicated zones based on the national platform, and support linkage between existing platforms and the national platform.

(15) Explore systems related to data rights and interests for AI development. In accordance with the principle of separating the three rights of data possession, data use, and data operation, clarify arrangements for the allocation of dataset property rights. On the premise of protecting lawful rights and interests, appropriately expand the boundaries of fair use under copyright. Balancing the needs of property rights protection and innovative development, explore a flexible regulatory system during the AI training stage, and create an institutional environment with clear rights and responsibilities that is compliant and inclusive.

(16) Persist in giving priority to ethics and in fairness and inclusiveness. Study and explore ethical norms for high-quality datasets, adhere to a value orientation based on benefit to society, and strictly prohibit the illegal collection or use of sensitive data. Guard against data bias and discrimination arising throughout the entire process

of dataset construction, fully consider the diverse needs of all sectors of society, and ensure to the greatest extent possible that the fruits of dataset construction benefit all of the people.

7. Implement the Initiative to Create Value

Give play to the application value of datasets, and use high-quality industrial datasets to empower AI development. Realize the value of datasets as a factor of production, promote dataset commercialization and assetization, cultivate a market consensus on paying for data, and explore a value system based on tokens.

(17) Give play to the application value of high-quality industrial datasets.

Establish a mechanism for linking dataset and model demand, promote precise adaptation between high-quality industrial datasets and models, improve model quality and efficiency, and deeply empower industry development. Promote the integrated use of high-quality industrial datasets across industries, fields, and scenarios, and encourage multiple application models such as “data-for-data exchange,” “data-model exchange,” “data trusteeship,” and “integrated data and compute.” Actively participate in open-source community development, encourage intermediaries and public-interest institutions to make foundational and public-interest datasets available to society as public goods, and stimulate vitality in data use.

(18) Innovate business models for high-quality industrial datasets. Improve long-term dataset operation mechanisms, encourage the listing of datasets for trading on data exchanges and other data circulation service institutions, develop diverse service forms such as the “subscription model,” “marketplace model,” and “customization model,” and promote the tier-by-tier upgrading of business models from the sale of basic data packages to application programming interface (API) calls, model-based solutions, and full-stack services. Explore new trading models such as token trading, and build a quantifiable and priceable dataset value system based on tokens.

(19) Explore innovative pathways for assetizing high-quality industrial datasets. Encourage qualified work units to take the lead in exploring pilot work on dataset asset inventory, registration, evaluation, and other activities, and accumulate replicable and scalable experience for data assetization. Encourage exploration of diversified innovative assetization models such as dataset pledge financing, exchange of appraised assets for equity shares (作价入股), asset securitization, data trusts, and data insurance, and broaden channels for converting the value of data.

(20) Cultivate a market consensus on paying for high-quality data. Establish and improve market-based benefit distribution mechanisms to ensure that entities at

all stages of the process, including data supply, processing, circulation, and application, can obtain market-based return on value and can share in data dividends. Encourage data demand-side entities to increase investment through purchases, cooperation, and other forms, give full play to the demonstration and guiding role of government departments, state-owned enterprises (SOEs), AI model enterprises, and other work units, promote the inclusion of data procurement in budget preparation, take the lead in launching data procurement practices, drive the formation of a market consensus on the paid use of data, and build a healthy and sustainable data market ecosystem.

8. Assurance Measures

The National Data Administration shall play the overall coordination role and, together with relevant departments, advance the construction and application of high-quality datasets in coordination, creating a work pattern featuring “departmental coordination and linkage across fields.” All local governments shall fulfill management responsibilities for their jurisdictions, intensify organization and implementation, formulate supporting measures based on actual conditions, and make overall arrangements for procurement funds for data products and services to support the building of high-quality industrial datasets. Support all types of entities in actively carrying out and participating in the construction of high-quality industrial datasets, guided by application. Channel financial institutions, patient capital, industrial funds, and others to increase investment in the construction of high-quality industrial datasets, encourage local governments to establish special funds, and explore diversified, multi-channel investment mechanisms. Establish and improve due diligence-based liability exemption (尽职免责) mechanisms, improve management systems for trial-and-error and error tolerance, and encourage innovative exploration on the premise of compliance with laws and regulations and keeping risk controllable. Continue tracking work on the construction of high-quality industrial datasets, improve monitoring indicators, evaluate the effectiveness of construction and application, and periodically summarize experience. Strengthen security assurance, implement the requirements of laws and regulations related to data security, establish a full-process security governance mechanism, guard against security risks such as data poisoning, data tampering, and data leaks, and firmly toe the bottom line for data security.