GOVERNMENT OF INDIA
MINISTRY OF ELECTRONICS AND INFORMATION TECHNOLOGY
RAJYA SABHA
STARRED QUESTION NO. *215
TO BE ANSWERED ON: 08.08.2025

STEPS TO DEVELOP AND PROMOTE INDIGENOUS AI DATASETS

***215. SMT. PRIYANKA CHATURVEDI:**

Will the Minister of ELECTRONICS AND INFORMATION TECHNOLOGY be pleased to state:

(a) the steps taken by Government to develop and promote indigenous Artificial Intelligence (AI) datasets which aid in the development of future AI platforms;
(b) the steps taken by Government to bring together academia, industry and experts to develop indigenous AI datasets;
(c) the steps taken by Government to make these datasets unbiased, fool-proof and vernacular; and
(d) the outcome in this regard?

ANSWER

MINISTER FOR ELECTRONICS AND INFORMATION TECHNOLOGY
(SHRI ASHWINI VAISHNAW)

(a) to (d): A Statement is laid on the Table of the House.

STATEMENT REFERRED TO IN THE REPLY TO RAJYA SABHA STARRED QUESTION NO. *215 FOR 08.08.2025 REGARDING "STEPS TO DEVELOP AND PROMOTE INDIGENOUS AI DATASETS"

……

(a) to (d): India's AI strategy is based on the Hon'ble Prime Minister's vision to democratize the use of technology. It aims to address India centric challenges, create economic and employment opportunities for all Indians.

AI ecosystem in India at present

- Fast-growing India's tech sector with annual revenue projected to surpass $280 billion this year
- 6 million+ people are employed in this sector
- 1,800+ Global Capability Centres, including 500+ focused on AI
- ~1.8 lakh startups in India; 89% of new startups in India last year were AI-powered
- Global rankings such as Stanford AI rankings placing India among the top countries in AI skills, capabilities, and policies to use AI
- India, second-largest contributor to GitHub AI projects, testament to its vibrant developer community.

Building on this strong foundation, the IndiaAI Mission was launched in 2024 to make AI accessible to all. It establishes a robust and inclusive AI ecosystem aligned with India's development goals.

Development of datasets for the development of AI is one of the main pillars of IndiaAI mission.

AIKosh – IndiaAI Datasets Platform

AIKosh is a unified data platform integrating datasets from government and non-government sources:

- Offers curated datasets across sectors such as health, agriculture, and education, with safeguards for data privacy
- Datasets are sourced from government departments, academic institutions, Indian startups, etc ensuring local relevance
- Available resources serve as building blocks for developers, allowing them to focus on core AI functionality instead of recreating modules
- India-specific 1200+ datasets and 217 AI models are available on the platform
- Examples of datasets - Farmer query data from Kisan Call Centres, geological data from states, clinical, imaging, & pathology data to support AI-based diagnosis of brain lesions
- Small AI models are also available on platform; For eg. Text-to-Speech (TTS) models in Indian languages like Bengali, Gujarati, Kannada, Malayalam
- Provides a sandbox mechanism which allows Indian start-ups/ academia to test their tools in a controlled environment

- Platform has attracted over 265,000 visits, 6,000 registered users, and 13,000+ resource downloads

Bharat Data Exchange (Bharat Data Platform) platform is an extension of Open Government Data (OGD).

- Serve as the data repository for AIKosh Platform
- Facilitate the access to Government owned shareable data and information in both human readable and machine readable forms

Digital India Bhashini

Bhashini is part of National Language Translation Mission (NLTM) which focuses on creating AI-driven language solutions.

- Citizens contribute voice, text, and translations in 22 Indian languages on **BhashaDaan** platform
- In collaboration with over 70+ research institutions & sectoral experts, large volumes of annotated datasets are curated for different technologies
- These include speech recognition, machine translation, & other language technologies
- MoUs with ministries, state governments, academic institutions, & industry partners to co-develop domain-specific datasets and enable cross-sector collaboration in Indic AI

Data is collected from people across different regions, communities, and backgrounds to reflect India's true language diversity and avoid bias. It includes real-life dialects and spoken variations, capturing the richness of India's linguistic landscape.

National Mission on Interdisciplinary Cyber-Physical Systems (NM-ICPS)

- **25 Technology Innovation Hubs (TIHs)** established in top academic institutions for domains like AI, ML, IoT, robotics, cybersecurity, and quantum tech
- **IIIT Hyderabad TIH**: Developed 105+ India-specific datasets (clinical, mobility, autonomous driving); digitised 2000+ pathology images, & developed India Driving Dataset (IDD) downloaded in 30+ countries
- **BharatGen consortium** (IIT Bombay, IIT Madras, IIT Kanpur, etc.): Built massive India-centric corpus—trillions of tokens, 1000s of multilingual speech hours, and millions of local documents.
    - The aim is to source diverse datasets and develop India-specific AI models
- **ARTPARK at IISc Bengaluru**:
    - Developed **Vaani dataset** (16,000 hours of audio across 54 languages, 80 districts)

- Developed (Medical-Imaging and Information Datasets) **MIDAS** medical imaging datasets for public health

Indian Council of Medical Research (ICMR)

- Health Research Data Repository for centralized, secure access to high-quality clinical datasets.
- Compliant with global standards (WHO, ISO, HL7) and national health protocols (NDHM/ABDM).
- Includes datasets from National NCD Monitoring Survey, ICMR-INDIAB study (2008–2020) – 113,043 participants
- TB treatment trials, diabetes registry, antimicrobial resistance network, and IN-CXR chest radiographs also included

IMPRINT + Uchhatar Avishkar Yojana (UAY): ₹1,000 crore allocated for AI curriculum, joint R&D, and industry-academia partnerships

Anusandhan National Research Foundation (ANRF) – "AI-for-Science" Initiative accelerates scientific discovery in physics, chemistry, and biology using machine learning models

"India AI Open Stack" initiative offers a foundational AI architecture embedded with science and engineering models tailored for Indian researchers

The outcome of these efforts is the development of high-quality, unbiased, and vernacular datasets that can be used for various AI applications, contributing to India's growth and development.

****