## AI Kosha: Democratising AI Innovation through Open-Source Datasets

Anupama Vijayakumar, Consultant, RIS

The Government of India recently launched AI Kosha, a platform consisting of non-personal datasets along with commissioning 14,000 GPUs for shared access. The two pillars are expected to support and foster innovation in artificial intelligence (AI) models suited to India's domestic context. Consisting of 316 datasets, the AI Kosha is expected to enable "the creation and validation of language translation tools for Indian languages, thereby promoting linguistic inclusivity and digital transformation". The platform consists of data ranging from the 2011 census to water quality, and the Defense Meteorological Satellite Program (DMSP) Night Lights Data (1992-2013). The "the bulk of these are datasets intended to help in creating or validating language translation tools for Indian languages". The AI Kosha would further gather data drawn from various gather non-personal data from various ministries and departments

India's approach to AI lays special emphasis on data requirements and standards for research and development in AI. The India AI mission launched in 2024 had laid stress on access to datasets as a key enabler of innovation in AI in addition to compute capacity, skilling and startup financing. One of the priority areas for the India AI mission was to make available diverse and high-quality datasets in various Indian languages to facilitate effective training of AI models. The focus on diversity and quality of datasets is also intended to ensure that AI systems remain free of bias and less prone to errors.

India has already set up the Open Governance Data platform (data.gov.in) which hosts 12,000 plus datasets. These datasets are further categorised as per sectors, institutions and states to the end "of encouraging cross-sector data sharing". Recent initiatives have also sought to curate domain-specific datasets. The Centralized Farmers Database which has been implemented under the National e-Governance Plan in Agriculture (NeGPA), for instance, seeks to develop a nationwide database of farmers for improving various activities "like issuing soil health cards, dissemination of crop advisories to the farmers, precision farming, crop insurance, settlement of compensation- claims, grant of agricultural subsidies, community/village resource centres etc".

Al Kosha can potentially boost India's efforts to help India develop cutting-edge home-grown Al models. As such, access to data is a major challenge for emerging and middle income countries to leverage opportunities offered by AI. This is due to the "dominance of the AI market by a few tech companies is due in part to their control over data that has been accumulated over time and scraped from the public internet". However, the advantage that India derives from its demographic dividend as well as linguistic diversity can help overcome such barriers and level the playing field. It further allows India to reduce reliance on Western data and bring it closer to achieving self-reliance in AI development and deployment.

Open source dataset platforms such as AI Kosha can enhance the discoverability and utility of data for supporting the use of AI for development. The platform is expected to boost the startup ecosystem and help AI work for India. It can truly help serve the needs of diverse groups and truly ensure "AI for All" in both spirit and substance. In the near future, the platform is expected to draw data from diverse sources to cover multiple domains relevant from a sustainable development perspective and covering multiple domains including health, education, agriculture and environment.

It further grants India credence as a global leader in AI. India can leverage the experience to build and foster engagements with the Global South in multilateral or minilateral forums. India can further look to build or improve these datasets to meet the UN's Digital Public Goods Standards which "operationalizes the UN's definition of digital public goods with a set of nine indicators, which ensure SDG-relevance, openness, fairness, safety and adherence to applicable laws". In doing so, India can take the lead in democratizing AI for the benefit of the Global South, in line with the ethos enshrined in *Vasudhaiva Kutumbakam*.