Ministry of Electronics and Information Technology
Government of India

National Language Translation Mission (NLTM)

Whitepaper

Contents

Executive summary

India is a vast nation consisting of multiple cultures, religions, diaspora and languages. Although only22 languages are recognized constitutionally, several hundred languages and dialects are spoken across the country. In the past decade, India has witnessed stupendous growth digitally - in 2019, the number of Smartphone users in rural areas surpassed that of urban India. The public sector has been a strong catalyst for India's rapid digitization. With the advent of the government's Digital India program, there has been a considerable demand for digital services. At the same time, private-sector innovation has helped bring internet-enabled services to millions of citizens and made online usage more accessible. This has led to a burgeoning market for digital products and services, going well beyond the borders of urban pockets.

However, due to the paucity of content in local languages on the internet, the Smartphone users are starved for content in languages they speak and understand. To support and cater to such a massive user base, it is required to have a robust content ecosystem, technological support for Indian languages, and access to open source data and translation tools.

At its very heart, language technologies are a powerful way to empower Indian language internet users as a demographic. Since most services are available online, it can positively impact the community by providing digital access to better services to speakers of that language.

The National Language Translation Mission (NLTM) has been announced in the budget by the Honorable Finance Minister in the backdrop of growing demand for accessing online services in local Indian languages. This will enable the wealth of governance-and-policy related knowledge on the Internet being made available in major Indian languages. The Ministry of Electronics and Information Technology (MeitY) has launched *"Bhashini"* to help ensure that digital content is readily available to all citizens, in their preferred languages.

The vision of Bhashini is to "*harness natural language technologies to enable a diverse ecosystem of contributors, partnering entities and citizens for the purpose of transcending language barriers, thereby ensuring digital inclusion and digital empowerment in an AatmaNirbhar Bharat* ".

The goal of Bhashini is to develop an ecosystem of innovative practices for data collection, curation, develop technology for speech to speech translation and deliver solutions powered by open data, apps and services. Bhashini shall act as an orchestrator to bring contributions (like data, models etc.) received from government, industry, academia and society into an open "Hundi" or "Repository". All contributions to Bhashini shall be validated and standardized using a

Unified Language Contribution API (ULCA). Bhashini shall ensure the development of open source datasets, tools, models, and technologies which shall be freely available to the ecosystem partners for innovation and research purposes. Bhashini shall also seek to provide computational resources to the start-ups for development of language oriented applications and services.

Bhashini shall set up a technological backbone containing the necessary software platform for data contribution, ingestion, processing, validation, tagging/labeling, model building, catalog publishing, and open access to data and model repository. It shall also include curation tools, benchmarking tools, analytics tools, dashboards, and leader boards and shall be hosted on the Bhashini cloud. There will be a reference app layer that shall contain applications that demonstrate innovative ways to leverage Bhashini. There shall also be an ecosystem apps layer that would consist of various private marketplace applications and solutions that utilize Bhashini's architecture and built for India to make discovery of apps and solutions easier.

Bhashini shall ensure the creation of large open source datasets and models by bringing all contributions into a shared repository. The contributions thus shared shall be validated and standardized using a Unified Language Contribution API (ULCA). Both individual contributions (through a crowd sourcing portal) and institutional contribution (directly via APIs) will be encouraged.

Bhashini shall run under the aegis of MeitY and shall have an Apex committee under the leadership of the MeitY secretary to set overall goals of Bhashini and review the progress from time to time. There shall also be an Executive committee, under the Apex Committee, to monitor the functioning of the various arms of Bhashini including foundation operations, research and ecosystem outreach. The TDIL Programme under MeitY will steer the activities of Bhashini and will also facilitate the Bhashini. It will also oversee the activities of all the verticals of Bhashini and will bind these towards the accomplishment of the objectives of Bhashini.

NHLT shall be the execution arm of the foundation layer and shall report to the Executive Committee via TDIL. NHLT shall be responsible for monitoring activities related to software engineering for the core layers, managing availability of computational resources, developing methodologies for data collection, validation and curation as well as manage all aspects of data and model processing within the repository.

The Ecosystem Engagement Unit (EEU) shall be required to manage ecosystem collaboration with state language missions and inter ministerial collaborations. The Ecosystem Engagement Unit shall also be responsible for organizing hackathons, monitoring research projects and

managing international collaborations and will also interface with Industry including MSME & start-ups.

Achieving the goals of Bhashini shall require contribution from all of the ecosystem players. The different stakeholders of Bhashini shall be as follows:

1. **Ministries / Departments**

   They shall empower Bhashini through identification of data resources that can be harvested and support Bhashini by aligning their language focused initiatives with that of Bhashini.

2. **State and Language Missions**

   State and language missions shall play a critical role in Bhashini. They will enable Bhashini through identification of usable data sources from state government entities& other sources and also drive crowd sourcing efforts through standard Bhashini tools.

3. **Startups**

   Startups will be encouraged to create innovative applications by using Bhashini's resources.

4. **Private companies with large digital reach**

   The industry players can assist Bhashini in multiple ways - building the data corpus via co-branded crowd sourcing initiatives, building the tech backbone of Bhashini, providing computational resources and volunteer talent for Bhashini. They can also contribute their open source data towards the common Repository of the mission.

5. **MSMEs**

   Indian MSMEs that may want to use data, NLP products and services of Bhashini to deliver innovative solutions to the market.

6. **Data collection and curation companies**

   They shall assist Bhashini through collection and validation of the dataset contributions using ULCA compliant tools.

7. **Publishers**

   A working relationship shall be established with book publishers who possess rights to books or any other literary works they produce, or may hold other rights on behalf of third parties to bring in large scale Indian language copyright free content without compromising their intellectual property.

8. **Citizens**

   Citizens shall be provided easy tools and motivated to contribute towards the languages of their liking through Bhashini's crowd sourcing platform. They will be the prime beneficiaries of this whole initiative.

Creating a multi stakeholder ecosystem for delivering a scalable solution for digitizing numerous Indian languages would require an evolving multi-year action plan. Bhashini's journey for the next three years shall consist of some key milestones that have been segregated across four tracks, namely Foundation Track, Contribution Track, Innovation Track and Grand Challenge Track. Each of these tracks shall be highlighted during the four quarters of the first year, however the preparatory work may start earlier and track will also continue beyond the quarter.

1. **Foundation Track (Highlighted in Q2 2021)**

   Bhashini will lay the foundation and focus on getting the data and contribution foundations ready, by publishing the ULCA API, rolling out the Data and the Model Repository and releasing a transparent and continuous benchmarking system.

2. **Contribution Track (Highlighted in Q3 2021)**

   The contribution will be kicked off after the foundation is in place. In this track, Bhashini will initiate contribution towards benchmark and training datasets in all languages across domains and spanning all language technologies. Bhashini shall activate data contributions from various government entities. Bhashini shall also activate the contribution ecosystem involving language missions, publishers, communities and citizens.

3. **Innovation Track (Highlighted in Q4 2021)**

   Once the contribution and foundation tracks are active, the innovation track shall be initiated. This track shall involve hackathons and challenge rounds for development of innovative applications, tools and platforms. Bhashini shall also develop inter-ministerial applications to leverage Bhashini's resources to provide citizen services. Bhashini shall also conduct workshops to encourage startups to utilize contributed data and models to build innovative solutions tailored to the needs of the Indian consumers, especially, the Bottom-of-Pyramid markets.

4. **Grand Challenge Track (Highlighted in Q1 2022)**

   In this track, which is initiated after having an initial set of open data and models in the repository, and having triggered the innovation cycle through hackathons and workshops,

Bhashini will launch a set of Grand Challenges with clear milestones and prizes. Mission will encourage academia, industry, start-upsand anyone from the society to rise up to participate in these grand challenges. Participants shall demonstrate their capabilities to fulfill the milestones set forth by Bhashini, which shall be aligned to the goals of Bhashini.

Given the ambitious scope of Bhashini, there is a need to unleash the full potential of the ecosystem, for which a clear strategy of orchestration is required. Hence, Bhashini has envisioned for itself an orchestration role to catalyze and energize a vibrant ecosystem that can self-sustain, amplify local communities, and enable decentralized innovation cycles. It now requires support from all the ecosystem players including the citizens. This would surely help the country usher into the next era of digital disruption.

1. India's digital leap

By many measures, India is already on its way to become a digitally advanced nation. India is home to one of the world's largest and most rapidly growing bases of digitally connected citizens and the rapid growth has been propelled by the public and private sector alike. It is digitising activities at a faster pace than many advanced and emerging economies.

The Digital India programme has resulted in a paradigm shift towards digitisation and e-governance in India. The programme has ensured the digital inclusion of all by providing access to robust digital infrastructure created under it, facilitating connection with the rest of the world. It weaves together a large number of ideas and thoughts into a single comprehensive vision to ensure that benefits of development reach each and every citizen of the country in equal measure along with faster and timely service delivery.

Focused effort towards developing the world's largest digital infrastructure to connect all the 2.5 lakh Gram Panchayats of India by optical fiber has now connected 1.54 lakh Gram Panchayats. Setting up 3.7 lakh Common Service Centres across India has not only encouraged digital entrepreneurship in rural areas but also improved access to digital services for common man. India has built a strong foundation of digital infrastructure and expanded internet access to around 76 crore citizens, with the world now looking at India as the one of the largest Internet user base with the lowest Internet tariff.

With 129 Crore Aadhaar holders, India is home to the largest population with a unique digital identity in the world. Digital services like e-Hospital, BHIM-UPI, online scholarships, DigiLocker, Umang app, e-Courts, Tele Law, eWay Bills etc. have improved ease of living for citizens. This is evidenced by 26.4 crore transactions over UPI worth Rs. 4.9 lakh crore in April 2021 and 7329 crore e-Transaction on National and State level e-Governance Projects in 2020 alone. The Government e-Marketplace (GeM) has not only made government procurement transparent but has also enabled small businesses and even startups to sell their products and services to government organisations. Resurgence of electronic manufacturing has made India the manufacturing location for the second largest number of mobile phones in the world.

These transformations speak volumes about the success of India. India is now poised for the next phase aimed at the creation of tremendous economic value and empowerment of lakhs of people as new age digital applications permeate and transform various sectors of the economy and governance.

1.1. Digitisation of the Indian economy & society

Public Digital Infrastructure such as India Stack has been a strong catalyst in the overall growth and transformational change in the economy. Digital platforms like Aadhaar, UPI, GeM, DIKSHA, DigiLocker and Co-win +have demonstrated the transformational potential of nationwide digital platforms. The Unified Payments Interface (UPI) platform has unleashed a revolution of no/low cost instant digital payments across the banking ecosystem. The DigiLocker platform permits users to issue and verify digital documents, obviating the need for paper. During the last year when schools were shut due to COVID-19, platforms such as DIKSHA enabled learning to continue for teachers and students across India. These platforms are driving India towards a future where access to knowledge, products, and services can be presenceless, paperless and cashless. Such platforms will strengthen access to knowledge, data governance, facilitate the use of new technologies such as AI, Blockchain, IoT and Geo-spatial technologies, and will significantly enhance the spectrum of services available online. As in the case of India Stack, these platforms will create opportunities for Indian startups and the industry to innovate and build globally competitive value-added services in multiple domains.

Digital India has brought a positive change in the life of Indians; the pandemic will significantly accelerate the shift to digital and fundamentally redefine the overall governance as well as business landscape in India. New business models and opportunities will continue to evolve and the digital transformation being accelerated will continue apace. Digital platforms are recasting the relationships between Governments, citizens and businesses and acting as catalysts to energise the Indian economy. Digital India can contribute significantly by growing capacities in digital technologies, and applying them across various sectors of the economy thereby forming the basis of our roadmap for unlocking the untapped potential of India's digital economy.

1.2. Public Digital Infrastructure

India is in a sharply accelerating "lift-off" phase of its digital journey. It is important to guide and promote the digital revolution by harnessing digital technology & fostering innovation. The government has been proactively promoting digitization to ensure that public services are made available electronically to citizens, even in the farthest corners of the country. The fact that many of these public digital platforms encourage open ecosystems allowed private sector innovation to bring internet enabled services to lakhs of citizens and made online usage more accessible.

1.3. India Stack

IndiaStack is a set of platforms and APIs that allows governments, businesses, startups and developers to utilise a unique digital Infrastructure to solve India's problems towards presenceless, paperless, and cashless service delivery. India Stack includes Aadhaar (for authentication and e-KYC), DigiLocker (for issue, storage and use of electronic credentials and documents), e-Sign (digital signature), Unified Payment Interface (UPI) and Aadhaar Enabled Payments (for financial transactions) and Data Empowerment and Protection Architecture (DEPA for privacy-protected personal data access and consented sharing). Together, it shall enable applications that could open up opportunities in the financial services, healthcare and education sectors of the Indian economy.

1.4. National Digital Educational Architecture (NDEAR)

The government has laid a major emphasis on strengthening the country's digital infrastructure for education by announcing setting up of a National Digital Educational Architecture (NDEAR) within the context of a Digital First Mindset where the digital architecture will not only support teaching and learning activities but also educational planning, governance administrative activities of the Centre and the States / Union Territories. It will provide diverse education ecosystem architecture for development of digital infrastructure, a federated but interoperable system that will ensure autonomy of all stakeholders, especially States and UTs. NDEAR builds on DIKSHA and enables foundation layers in the education domain along with both learning and administrative interaction services.

1.5. National Digital Health Mission (NDHM)

The National Digital Health Mission (NDHM) aims to develop the backbone necessary to support the integrated digital health infrastructure of the country. NDHM shall create a seamless online platform "through the provision of a wide-range of data, information and infrastructure services, duly leveraging open, interoperable, standards-based digital systems" while ensuring the security, confidentiality and privacy of health-related personal information.

The government accelerated the digitisation process by building a foundation of digital infrastructure and public platforms, then introducing digital applications and services, and encouraging the ecosystem to innovate around it. These created real incentives for citizens to go online. However, the majority of the information available across all

platforms is in English, thereby hampering the public reach and accessibility to these services.

2. Need for a public digital Infrastructure for Indian languages

India is a country with great linguistic diversity. Even if we only consider the official languages as declared by the Indian constitution, we have 22 languages, each of them being spoken by over ten lakh people. Over the last decade, the enormous penetration of the Internet in the country means that a major chunk of this multilingual population is actively seeking to consume and interact with online content in their own languages.

2.1. Paucity of Indian language content

India has over 75 crore people using the internet, not particularly in English, but in their own languages as well. Indian users use the internet across low friction verticals such as entertainment, news, messaging, and social media. The number of Indians using the internet for utilities such as ticket booking, calculators, calendars, and notes has also been increasing steadily.

However, the Internet today is severely deficient in terms of content for Indian languages. In fact, 53% of non-Internet owners in India state that they would start using the Internet if it has content available in their local languages. The lack of local language content is particularly stark in the multimedia domain that consists of videos, podcasts and digital assistants.

2.2. Unavailability of resources in native language

Although a sizable user base is already connected with the internet and actively online, various barriers exist that prevent these users from engaging with more verticals and availing more services, and truly enjoying what the internet has to offer them.

- Much of the information (and therefore access to internet, e-governance, e-commerce, e-banking etc.) cannot be used by the majority of the population as it is available majorly in English. Unfamiliarity with English means that Indian language users hesitate using these services, as they don't completely understand what is being offered to them. They feel safer using services in their own language, even if that means availing these services in person.

- Text information is majorly available in English only except for a few languages (e.g. text books, information on government schemes, crop advisory etc.) and due to lack of English reading/writing skills a large percentage of the population is not able to tap the complete benefits of the internet revolution and is also unable to avail the full benefits of government schemes.

- Content creation poses another major problem, with most users (who use mobile devices) finding friction points when it comes to typing and viewing Indian text.

Therefore, in India's new and emerging digital ecosystems of the future, we must find novel ways to engage with citizens who communicate in varied languages and dialects. The objective is to make information available to the people in their native language in order to be a "truly" connected nation.

It is essential that a shared public digital infrastructure be created to open up technologies, data, and reference models so that everyone in the ecosystem can rapidly innovate on top of the public infrastructure thus catalyzing creation of a large Indian language knowledge base and availability of diverse solutions across all domains.

3. Bhashini

3.1. Background

With heightened Smartphone proliferation, availability of cheap mobile data, expanding WiFi services in villages and overall digital literacy, India has an unprecedented opportunity to create a blueprint for building the Internet for local languages. Bhashini, also known as the National Language Translation Mission (NLTM), has been envisaged to take advantage of this opportunity. Bhashini aims to build a distinct Indian language technology platform, related services and products, by leveraging the power of artificial intelligence.

Over the last decade, technology development has been done in the areas of automatic speech recognition, text to speech synthesis, machine translation and optical character recognition under TDIL Programme using Rule Base, Statistical Techniques and using various open source engines etc.

- The Automatic Speech Recognition (ASR) software for Indian English and 11 Indian Languages namely Hindi, Bengali, Marathi, Tamil, Telugu, Malayalam, Gujarati, Odia, Assamese, Manipuri & Kannada have been developed using Kaldi Open Source Speech Recognition Toolkit. Limited vocabulary continuous speech recognition

(LVCSR) Mandi system has been developed which provides speech-based access to agricultural commodity prices and weather information in these 11 Indian languages/dialects.

- Text-To-Speech (TTS) synthesis system developed is leveraged to send the information retrieved from Mandi System through voice SMS as well. The Browser TTS plug-ins also help visually challenged in accessing digital information. The TTS system (m-Vachak) has also been made available on Android based operating system, Indus OS and as of now there are 1.54 million activations on 8 Mobile Brands (Micromax, Celkon, Swipe, Karbonn, Intex, Trio, Sansui & Datawind) supporting Indus OS.

- Machine Translation (MT) Engines have been developed to connect English with other Indian Languages and also other Indian Languages(ILs) with Hindi.
  English to Indian Language MT: Angla MTSystem**,** a rule based machine translation system for English to 8 Indian languages [Hindi, Malayalam, Bengali, Urdu, Punjabi, Tamil, Assamese, Nepali]**;** Anuvadaksh System, statistical and Example based for translation from English to 8 Indian languages [Hindi, Marathi, Bengali, Urdu, Odia, Tamil, Gujarati, Bodo].

- Indian Language to Indian Language Machine Translation (ILMT) Sampark System: It uses both traditional rules-based and dictionary-based algorithms with statistical machine learning approach developed for translation of 7 languages to Hindi [Punjabi, Urdu, Tamil, Telugu , Bengali, Marathi, Kannada] In addition, it also addresses translation among Dravidian languages viz. Telugu <-> Tamil, Malayalam <-> Tamil.

- The OCR system has been developed for 13 Indian languages/scripts- Assamese, Bengali, Gurmukhi, Hindi, Kannada, Malayalam, Tamil, Telugu, Urdu, Gujarati, Oriya, Manipuri and Marathi, which can be used under Windows/Linux environment. The Web version has also been developed with Post-processing feature.

- Basic Information processing Kits across 22 scheduled languages have also been made available in open source.

- A large number of tools such as taggers, spell checkers, CLDR, grammar checkers, sorting utilities etc. have been developed for all major Indian languages which have

been used in several applications for Indian languages. A large repository of linguistic resources like corpus, wordnet, dictionaries, thesauri, tagged lexicons etc. has also been developed.

In addition, since the year 2000, active participation in global Standardisation initiatives like Unicode, W3C, ISO etc. facilitated incorporation of Indian Language requirements into these standards. Notable contributions are interoperability of Indian languages data across platforms, browsers and applications, the standard Indic Syllable definition solved long pending issue of seamless rendering of Indian Languages data using open font format standards across.

A significant stride was taken by setting up TDIL data Centre for proliferation of the research in this area further by propagating the use of free annotated linguistic resources by Researchers and Research Scholars towards this national cause. Subsequently, NPLT Platform was set up to disseminate the above outcomes to engage the stakeholders like start-ups, MSME etc. The goal is to enhance the ecosystem and play a proactive role in proliferation of language technology. The quality of research also needs to be constantly monitored to reach the desired outcome.

Globally speech recognition, machine translation & OCR technologies have been making dramatic improvements due to data driven Artificial Intelligence (AI), specifically deep learning technologies. Hence the research efforts need to be re-oriented to leverage AI and deep learning technologies. However, the key factor in building solutions using deep learning is the accessibility to multilingual data. India can tap a large amount of data available across Indian languages but in its current form it is unusable i.e. progress in Indian languages is hampered due to the lack of availability of large amounts of high quality datasets to train state-of-the-art AI models. Hence, it is required that curated and validated contributions should flow from varied sources, into an open source repository where it can be harnessed and put to use by the entire ecosystem.

A major step in this direction was taken by initiating Mission-Pilot Projects under which AI and deep learning techniques have been leveraged to establish feasibility of Speech-to-speech technology development with injunction of linguistic features based on the prior research expertise. These encouraging results lead to the conceptualisation and formulation of Mission Bhashini.

Bhashini intends to build advanced open source datasets and models in language technologies. These will cover the areas of Input Tools, Machine Translation (MT), Automatic Speech Recognition (ASR), Text-to-Speech (TTS) and Optical Character Recognition (OCR). In addition, the project will also focus on building datasets and models for natural language understanding (NLU) tasks such as Named Entity Recognition (NER), Sentiment Analysis, Question Answering (QA), and Summarization in Indian Languages etc.

The aim is to apply these developments to all Indian languages viz. beyond the 22 official languages as well. There should be a special focus on the low resource languages of the North-East and tribal areas. This would require a concentrated and coordinated effort that would include State government and local institutions, startups, researchers, scientists, academia and industry. The expansive cross sector participation will encourage collection of data, knowledge sharing and innovative practices to solve particular use cases.

Bhashini shall be carried out under the aegis of the Ministry of Electronics & Information Technology (MeitY). MeitY has already been supporting research activities in the area of language technology under the Technology Development for Indian Languages (TDIL) program. MeitY will leverage these research initiatives and the expertise offered by the EkStep Foundation to implement this Mission. It will also catalyse Bhashini's activities by aligning its efforts with other inter-ministerial missions by coordinating it with the Department of Industrial Policy and Promotion (DIPP) and PM-STIAC Secretariat viz. Invest India for support in outreach activities including State Government Missions.

3.2. Vision

The vision for Bhashini is to:
"*Harness natural language technologies to enable a diverse ecosystem of contributors, partnering entities and citizens for the purpose of transcending language barriers, thereby ensuring digital inclusion and digital empowerment in an AatmaNirbhar Bharat*"

3.3. Mission

The mission of Bhashini is to:
"*Create a knowledge-based society by transcending the language barriers and providing content and services to citizens, in their own language, both in the form of speech and text.*"

3.4. Objectives

1. To build a high-quality speech to speech machine translation (SSMT) system for major Indian languages.
2. To create and nurture an ecosystem involving start-ups, Central/State government agencies working together to develop and deploy innovative products and services in Indian languages
3. To increase the content in Indian languages on the Internet substantially in the domains of public interest, particularly, governance-and-policy, science & technology, etc

3.5. Purpose & scope

One of the main purposes of Bhashini is to develop a national digital public platform for language to provide universal access to content i.e. boost the delivery of digital content in all Indian languages. This would result in the creation of a knowledge based society where information is freely and readily available and would make the ecosystem and citizens “AatmaNirbhar”.

Bhashini shall also strive towards development of an ecosystem of innovative applications and citizen services that leverage the open datasets and models created by stakeholders of Bhashini. This will ensure the creation of multilingual content and hence provide an impetus in preserving Indian languages in this digital era.

Thus Bhashini shall act as an orchestrator to unify and align a large diverse network across government, startups, society and corporate entities and bring all their contributions (including data, models etc.) into an open repository.

3.6. Guiding principles

- **Bhashini shall be an orchestrator to catalyze a collaborative ecosystem**

  Bhashini shall create and nurture an ecosystem involving government institutions, industry players, research groups, academia, individuals etc. to develop an ever evolving repository of data, training and benchmark datasets, open models, tools and technologies. Crowd sourcing initiatives shall be set up to engage citizens and local communities directly for creation of original and translated content.

- **Bhashini shall enable interoperable, organized open data and model repositories through its open source foundation layer and a set of unified APIs**

Bhashini shall ensure the creation of large open source datasets and models by bringing all contributions (both institutional and citizen) into a shared repository that is fully open for innovation. All such contributions shall be managed using a Unified Language Contribution API (ULCA).

Bhashini shall ensure that the datasets, models, technologies, tools, and all other components developed during the project shall be open source and freely available to all contributors and end users. The quality of research shall be constantly monitored to reach the desired outcome. The easy availability of structured and organized data, reference models, and tools in open source shall make available opportunities for rapid innovation by the ecosystem.

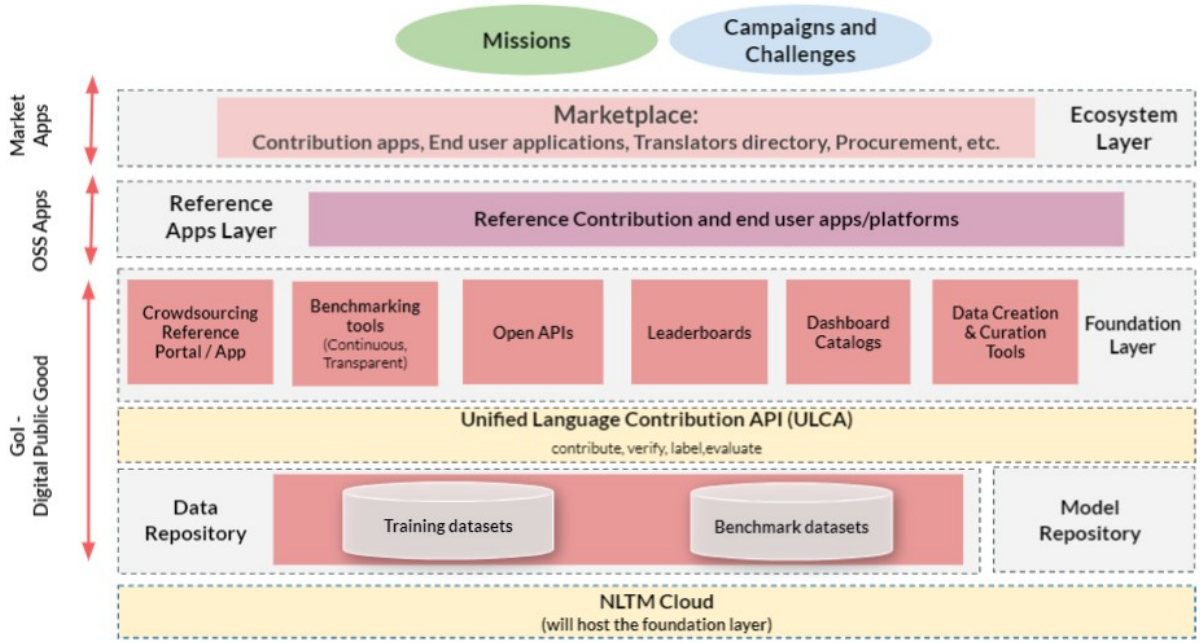
- **Bhashini shall strive towards solving real world problems by supporting innovation**

Bhashini shall encourage the ecosystem to develop innovative products and services for the citizens by leveraging the open repository of datasets and models. This shall include ensuring availability of computational resources to developers and researchers and exploring innovative ways to reduce the cost of innovation and improve the quality of data collection and curation in collaboration with the ecosystem players.

Bhashini shall also conduct challenge rounds and hackathons, inviting participation from individuals, industry, and academia, to think up novel solutions in the language technology domain.

4. Creating a unifying architecture

Bhashini shall create a unifying architecture, underpinned by principles of open data and open source software, to enable contributions from the research initiatives and the ecosystem. This shall also catalyze the ecosystem to work on an integrated approach to build diverse solutions on top. The idea is to build a community of contributors that works with a unified approach to help Bhashini realize its stated objectives.



The architecture of Bhashini shall follow a multi-layered approach. The different components of the architecture are as follows:

4.1. Bhashini cloud

The Bhashini cloud is where the different components of the foundation layer and data repository will be hosted including the following components:

- Data repository
    - Training datasets (MT, ASR, TTS, OCR & other NLP related datasets)
    - Benchmark datasets (MT, ASR, TTS, OCR & NLP related datasets)
- Model repository
    - Open source models (MT, ASR, TTS, OCR & other NLP models)
- Foundation layer applications
    - Data validation and curation tools and utilities
    - Benchmarking tools
    - Open API

- Leader boards, Catalogs, Discovery tools
- Reference Crowd sourcing Application

The Bhashini cloud shall consist of the following parts:

- Host applications and data in the foundation layer of Bhashini.

- A significant amount of GPU compute is required to train language AI models. High end GPU resources from CDAC shall be utilized for open source model training by the model contributors. GPU based model inference services (hosted at CDAC) to be consumed by applications built by the ecosystem. Hardware resources from the National Supercomputing Mission may also be leveraged for training of models.

- Provide high compute cloud credits from ecosystem partners for startups working in the language technology space as an enabler.

Bhashini cloud shall have a cloud agnostic policy which will embody the following principles:

- It shall be vendor neutral for open source, proprietary, public, private, hybrid cloud deployments both in short term and long term. There are many proprietary cloud stacks offerings that are available as both public, private and hybrid versions in the marketplace. Open stack is a popular open source cloud stack that many organisations use to host a private cloud

- All applications shall be capable of seamlessly running on any type of cloud hosting (i.e private, public, hybrid using open-source or proprietary cloud stack). The key to achieve this principle is to abstract the key aspects of the infrastructure (i,e compute, storage, runtime, etc.,) to well known open standards. Application deployment should be automated such that the same application code-base should be able to run on multiple clouds setups. For e.g., Development/Test instance can be Public Cloud 1, Benchmarking instance on Public Cloud 2 and Production instance on private hosted cloud.

These principles should be followed by all public digital infrastructures and it is recommended that a separate policy should be created for the same.

4.2. Data Repository

The data repository contains training and benchmark datasets. It is the vision of Bhashini that this will be the largest repository of data to drive language technologies in Indian languages. Datasets in the repository would be ingested, validated, labelled, and curated using the Unified Language Contribution API (ULCA) backbone. Reference open source tools and utilities will be used for data ingestion, validation and curation.

The Data Repository shall ingest data from varied sources, including:

- Public (non-personal, open) data from data sources such as government agencies (Prasar Bharati, NCERT, DAVP etc.)

- Scraped data from the web, processed into a usable form.

- Data contributed from previous language projects sponsored by MeitY (from Academia, TDIL, CDAC etc.) and CIIL, Mysore, MHRD, Department of Official Languages, MHA etc. Available open source corpora in Indian languages. Also contributions from the corporate sector will be explored.

- Language data in Indian languages collected by NGOs and research foundations

- Crowd sourced data received from the public at large collected through various channels including those of industry partners

- Gold standard human curated data which would be collated under Bhashini on a regular basis.

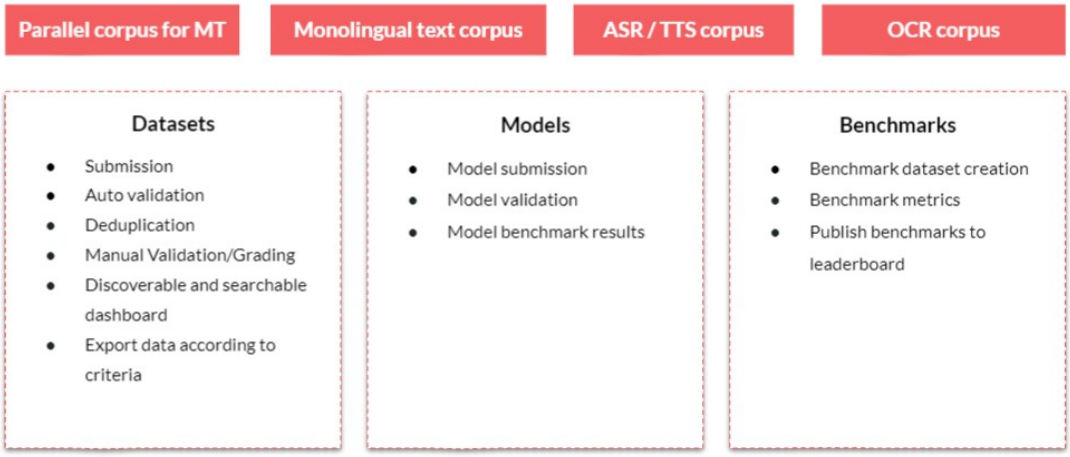
4.3. Model Repository

Bhashini will aim to create state of the art open source models (MT, OCR, ASR, TTS) for Indian languages. The model contributions shall come from research groups, startups and volunteers working on language technologies. The contributors shall be encouraged to share models across languages and domains so as to create a comprehensive open source model repository.

4.4. Unified Language Contribution API (ULCA)

Bhashini shall capture all data and model contributions through the Unified Language Contribution API (ULCA). ULCA will standardize all data and model contributions for benchmarking. This is critical to remove data silos, duplication and ensure the dataset discoverability and quality.

Efforts shall be made to reduce the cost of data collection through innovative methods such as crowd sourcing and scraping of data on the web. All data contributed to Bhashini through ULCA shall be attributed to the original contributor / source.



4.5. Foundation layer

The foundation layer of Bhashini consists of the set applications and utilities which will enable orchestration among the ecosystem partners. It consists of the following major components:

4.5.1. Crowd sourcing reference application

Crowd sourcing tools will be created and promoted for the purpose of Indian language data collection from the general public. This will allow local communities and individuals from across India to contribute to Bhashini. Bhashini will develop reference crowd sourcing applications (Bolo India, Likho India, Padho India etc.) which will be integrated to the data repository through ULCA.

In addition, ecosystem players will be actively encouraged to develop their own crowd sourcing applications and other community engagement apps to allow decentralized and customized data collection (with Bhashini branding) and then contribute back to the data repository in bulk via ULCA.

4.5.2. Data creation and curation tools

Data creation and curation tools will be integrated with the data repository through ULCA. Pipelines will be developed to collect, validate, tag and curate various kinds of data including web scraped and aligned data, crowd sourced data and human curated (benchmark data) to ensure the quality and usability of the data collected.

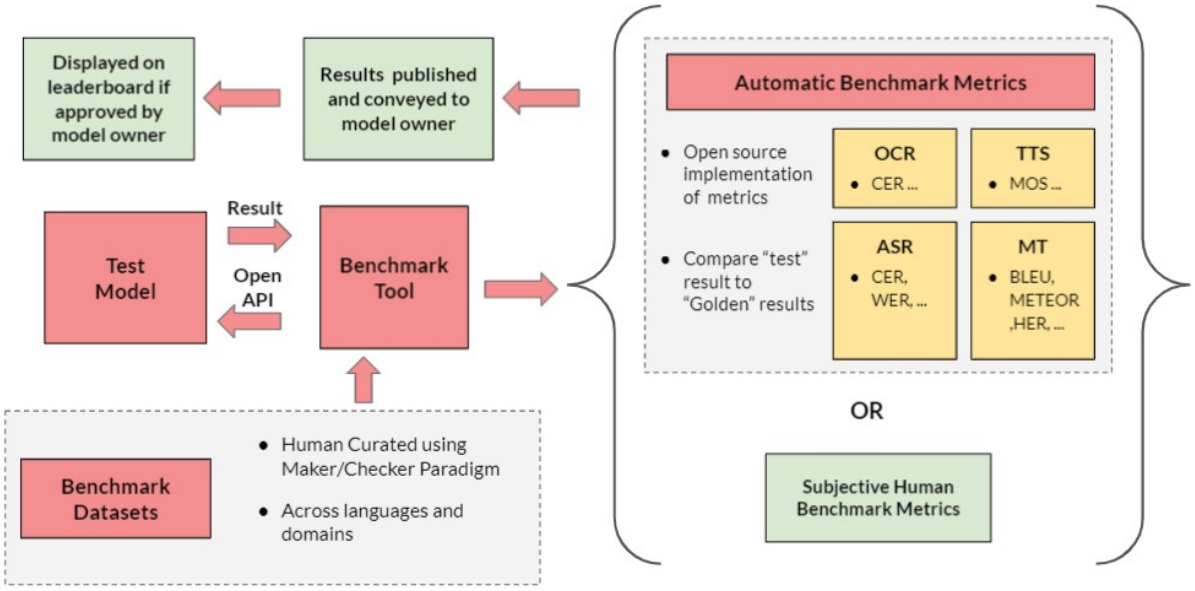
4.5.3. Open API

Any model whether open source or not can be benchmarked using the benchmarking system as long as it implements a REST service compliant with the ULCA API. This API is also the interface to access the model by applications.

4.5.4. Benchmarking tools

Benchmarking is a critical activity for the progress of language AI. A transparent and continuous benchmarking system will be established to test all models.

Benchmark datasets shall be extensive and come from different sources (domain, source, speaker, font etc.) as diversity is important for robustness. Development of benchmark datasets shall be an ongoing process to ensure freshness and shall employ a maker-checker paradigm for data check. The process for benchmarking shall be as follows:



- Datasets shall be ingested and then validated using an Open API defined by ULCA to create benchmark datasets. The curation shall be manual using multiple human curators. The benchmark datasets thus prepared shall span across varied languages and domains, These shall be utilized by the benchmark tool for validating models submitted by contributors.

- When a test model is submitted, the benchmark tool shall be called using an open API. Every model that is submitted will automatically run on all relevant benchmarks using the benchmarking tool. The benchmark tool shall evaluate the test model on the defined benchmark metrics. Benchmarking metrics shall be agreed upon by the entire ecosystem to measure the performance of systems.

- Results obtained from the benchmarking exercise shall be published and conveyed to the model owner. If approved by the model owner, the benchmark results shall be published to a leader board.

4.6. Data Sharing Policy

In the age of AI and deep learning, large amounts of data are the key to building high quality language models. While commercial entities have access to tremendous amounts of public and non-public data that they utilize for building models, the public initiatives suffer from a paucity of data. Therefore, it is the important for Bhashini's data sharing policy to level the playing field in the context of Indian Languages

The Ministry of Science & Technology (MoST) has notified the National Data Sharing and Accessibility Policy (NDSAP) for open sharing of data created using public funds that are non-sensitive or non-personal (not explicitly classified as non-sharable). The policy shall apply to all data and information created, generated, collected and archived using public funds provided by the Government of India directly or through authorized agencies by various Ministries, Departments, Organizations, Agencies and Autonomous Bodies. The aim of the policy is to promote data sharing and enable access to government owned data for national planning and development.

A linguistic data policy will be formulated for NLTM and other such initiatives.

4.7. Reference apps layer

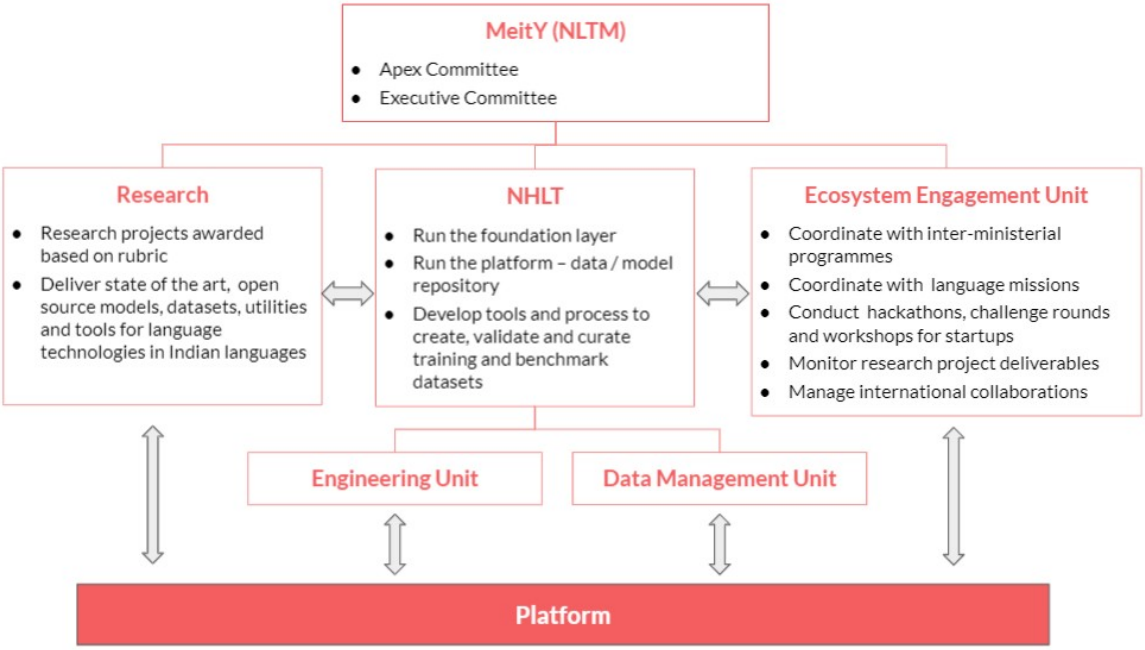
The MT, ASR, TTS and OCR models developed in the foundation layer can be used to create applications in various domains. A set of open source reference applications may be developed to demonstrate innovative ways to leverage Bhashini's resources. Applications may use the models on an as-is basis or might fine tune the models to target specific applications. Reference applications should also implement an ULCA interface to help contribute post-editing results and statistics back into the data repository.

4.8. Ecosystem Apps Layer

The ecosystem apps layer consists of the tools and applications available in the market that leverage the Bhashini architecture. These applications would primarily be developed by the industry including startups. These applications are not expected to be open source. The ecosystem apps layer will also include missions that are developed in collaboration with other ministries. Examples of such applications include:

- A translation system to effectively translate government documents into 22 Indian languages while minimizing the turnaround time and human effort involved.
- A system to evaluate the oral reading fluency (ORF) for primary school children in their mother tongue.
- A voice based mobile payment application to allow all Indians to make secure payments using their own language.

- Indian language speech and text chatbots, digital assistants etc. to spread awareness and better access to citizen centric schemes of Central and State Governments



5. Governance Structure

5.1. Ministry of Electronics & Information Technology (MeitY)

MeitY has been the driving force behind innovations related to language technology in the country. Through it's TDIL program, MeitY has taken great strides to empower citizens in the digital space. By means of language processing tools various digital applications are now being made available to the masses of India.

It is proposed that the project be funded by MeitY with a goal to overcome the language barrier among all major languages of India in speech and text. The efforts would be made in the direction of funding corpus development so that the Indian industry partners may also contribute to the cause in a like for like manner, by offering technical resources, open source data, applications, compute or storage for supporting the initiative. The TDIL Programme under Meity will steer the activities of Bhashini and will also facilitate the

smooth implementation of Bhashini. It will also oversee the activities of all the verticals of Bhashini and will bind these towards the accomplishment of the objectives of Bhashini.

Bhashini shall have an Apex committee under the leadership of the MeitY secretary to set overall goals of Bhashini and review the progress from time to time. Bhashini shall have an Executive committee to monitor the functioning of the various arms of Bhashini including foundation operations, research and ecosystem outreach.

Bhashini shall have three entities namely, Research, National Hub for Language Technology (NHLT) and Ecosystem Engagement Unit (EEU), working on different objectives. The two verticals of NHLT and EEU will be implemented through two separate agencies, looking at the diverse nature of activities of the two verticals. However, the entities shall work in cohesion, under the direction of MeitY, for fulfilling their stated objectives.

5.2. National Hub for Language Technology (NHLT)

NHLT is the body which is responsible for engineering, reference application and data management. To achieve this, NHLT will consist of an Engineering Unit and a Data Management Unit.

NHLT will develop the platform leveraging but not limited to open source components developed by ekStep foundation. The platform will host the open data repository along with other sub activities mentioned below:

- Open data repository, Model repository and catalog.
- Benchmarking tools
- Leader boards
- Hosting and sharing API as service as required
- Data collection and curation tools
- Dashboard for startup support including credits and API allotments.
- Develop and maintain Bhashini portal.

Karya, an open-source platform from Microsoft, will be leveraged to democratize various language data preparation tasks such as data sourcing, data validation, data annotation and labeling, transcription etc. This would be continuously enhanced in collaboration with contributors of the Karya open source system.

Allocation of compute resources to various NLP / AI start-ups / researchers / research institutes / innovators etc. will be managed by NHLT. NHLT will further also manage quota for free API uses by startups and MSME.

NHLT will be advised by three expert groups for consultation on data standards and quality in the field of speech, text & OCR. The group can consist of national or international technology experts, industry members and linguists.

NHLT will develop various training programs for data preparation tasks by providing all infrastructural support including hardware and engineering.

5.2.1. Engineering Unit

The Engineering Unit shall be responsible for operationalising and maintaining the foundation of the Bhashini platform. The Engineering Unit will carry out the below mentioned activities through empanelling of Software Development Agencies (SDAs) for providing engineering support to Bhashini and also building an in-house team. The Bhashini cloud architecture should be designed in a manner that vendor lock-in is avoided and it should be easily scalable and portable across public-private cloud infrastructure as per Meity empanelment. The government cloud infrastructure from C-DAC and NIC may be leveraged to the extent possible.

- Operationalization and maintenance of the Bhashini platform
- Run the data and model repository through the ULCA interface
- Run the foundation layer of the Bhashini Cloud
- Benchmark contributed models using the benchmarking tools
- Manage data and compute resources for language technology startups and researchers

5.2.2. Data Management Unit

The Data management Unit (DMU) shall be responsible to create an efficient pipeline to manage the data collection and curation process compliant to ULCA. It shall perform the following functions:

- DMU will set up data collection protocol, development of data collection/validation/scraping tools and workbench. To define data standards the DMU will further consult with the expert groups formed by NHLT on Speech, Text and OCR. All the data collection protocol updates would be defined by DMU for Bhashini and would be responsible for data management.

- DMU will further take requests for data creation from research proposals and other stakeholders of Bhashini and will float requirements on GEM (or other e-Market Place) for fair bidding in terms of multiple bundles for creation of said data . DMU will further post and manage data creation of any type in Bhashini.

- On approval of a new data collection requirement, the proposing entity will define the data collection/validation and curation task as a template in the data collection pipeline (or use an existing pipeline). This definition should be reviewed to ensure best practices. They will also provide all the technical details of the quality of the data to be created. This task may be divided into smaller subtasks to allow tighter control on data quality by DMU. For example, a task to collect 50,000 benchmark sentences (Eng-Malayalam) may be broken into 10 tasks for 5,000 each containing both a “maker” and “checker” entity and also various automated checks.

- Each task would then be made available to all empanelled agencies which are qualified to do that task. The agency shall be selected in a transparent manner (QCBS) and the task would be assigned. A contact person in the proposing entity along with DMU would coordinate with the agency. Upon completion of each task the validation process shall be done and a reputation score of the data collecting agency would be updated. DMU will also enable empanelment, rating etc of data procurement and validation agencies in the e-Marketplace

- DMU will ensure that pipelines should be built to collect data for input tools, MT, ASR, TTS, OCR. In addition, pipelines may be developed for other NLU tasks such as Named Entity Recognition (NER), Sentiment Analysis, Question Answering (QA), Summarization etc. Pipeline should be developed to collect data from different categories including the following:

    - Pipelines for data scraped from the web and aligned using automated methods. This data is of high volume but may be noisy. Such datasets have proven to be effective for pre-training large scale AI models.
    - Data collected through crowd sourcing approaches with automated techniques to validate quality. This data is medium volume and of higher quality than the scraped models. It may be used during the fine-tuning of models for a specific language or domain
    - Data that is carefully curated by experts. This data has low volumes but is of the highest quality. This should be for benchmarking purposes.

- The Unit will also build pipelines for the validation of each dataset. Each dataset contribution shall be independently validated through a combination of automated tools and human validators to evaluate the quality of data submitted. The validation process shall be monitored and a sampling process may optionally be used to determine data quality.

- The team will ensure transparency in the discovery of datasets. Dataset information including quality attributes shall be reflected in a dashboard. Users may download datasets that meet their training needs from the repository based on their requirements. Actual data collection will be done by empanelled data curation agencies and freelancers, using the tools and methodologies specified by the Data Management Unit. The DMU shall monitor quality levels in data and later maintenance of data quality as feedback is received from users. Bhashini will study similar setups such as LDCIL in Mysore, LDC in USA, and ELRA/ELRC in Europe to ensure that best practices are followed.

- The Data Management Unit will also establish open source model training pipelines to train open source models for ASR, TTS, OCR, MT and various NLP tasks in various languages. These models will be benchmarked to establish baseline performance.

- DMU will further manage ULCA compliance of data for interoperability of data and will work closely with partner institutes, open source groups, State Anchor Points (SAPs), crowd-sourcing agencies. To execute the operation the DMU may partner with research groups apart from in-house capacity.

- DMU may develop in-house resources if required for certain low resource language or specific tasks for quality control. State language missions will also provide resources to DMU for all languages.

5.3. Academia & Research groups

Academia and research groups shall be responsible for carrying research in the areas of language technologies and translation. Since fundamental research with a long term horizon is important for the success of Bhashini, the research groups will be required to submit their proposals for the same, clearly outlining all objectives and deliverables. The projects will have short term (quarterly) and medium term (yearly) deliverables.

Bhashini shall draw up a rubric for evaluating the research proposals as medium term or long term. All research proposals shall be evaluated by an expert committee against this rubric in a fair and transparent manner.

Research proposals may often require data collection. The nature and amount of data to be collected should be justified taking into account the state of the art and available datasets. It is possible that the amount or nature of data to be collected changes due to technological trends. All data collected must be compliant with ULCA processes identified by the Data Management Unit to allow sharing and reuse of data with the ecosystem at the earliest possible following standard process.

Models developed as part of the Government funded research projects must aim to improve open source baseline benchmark results which will be updated from time to time to reflect the state of the art.

Bhashini shall also allow submission of proposals for short term projects, as the same may be required to make rapid progress on Bhashini. Bhashini shall decide the relative allocation of short, medium and long term projects for research. Central and state funded institutes' led consortia shall also be allowed to participate.

5.4. Ecosystem Engagement Unit

Bhashini will have an Ecosystem Engagement Unit (EEU) to manage ecosystem collaboration with state/ language missions and inter ministerial collaborations. The Ecosystem Engagement Unit will be responsible for ecosystem outreach in Bhashini. The Ecosystem Engagement Unit will liaison with the state language missions to create training data and develop content in regional languages. Ecosystem Engagement Unit will also liaison with various Central Ministries to develop inter ministerial missions leveraging the resources of central ministries. The Ecosystem Engagement Unit on behalf of Bhashini may seek the services of professional media agencies to assist the government in public engagements and branding of government activities.

The Ecosystem Engagement Unit will be responsible for organizing hackathons and challenges in Indian language technology and encourage the participation of startups to develop applications (or models/datasets/utilities) as part of Bhashini. The Ecosystem Engagement Unit shall also be responsible for monitoring the deliverables of the research projects and managing international collaborations for propagating Bhashini. It shall be responsible for the following major activities:

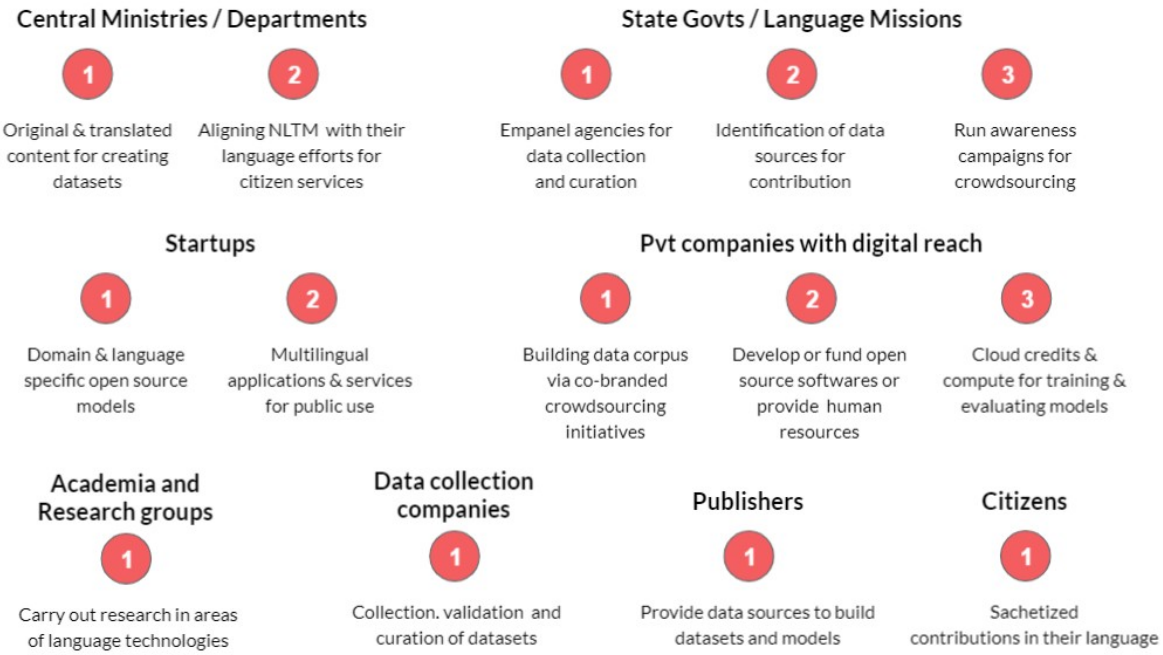
- Assist MeitY in monitoring of the activities of Bhashini across all verticals
- Set up meetings of Apex Committee and Executive Committee in coordination with Meity
- Prepare suitable Progress Reporting Formats and mechanisms for Research Projects
- Evolve mechanisms for reporting the benefits accrued out of startup engagement
- Prepare Inputs on Mission as required by Meity for reporting progress of Bhashini to PMO (Prime Minister Office), e-Samiksha, Sectoral Group of Secretaries(SGOS), OOMF (Object Output Monitoring Framework)
- Establish international linkages for replicating success stories
- Unifying state language missions across states and union territories

6. Catalyzing the ecosystem

Bhashini shall strive towards aligning all the contributors of the ecosystem consistent with the Bhashini architecture. At the outset, a contribution narrative shall be prepared for igniting

collaborations with ministries, industry leaders, and startups, large corporations including MNCs, philanthropists and NGOs. The primary support that would be required from all these players would be for data crowd sourcing and provision of unbundled contributables.

Some of the activities and objectives of the major stakeholders are outlined below:



6.1. Central Ministries / Departments

The central ministries and state departments can offer their contribution to Bhashini through identification of original and translated content through sources such as national and state broadcasters, examination boards etc. This data will need to be processed into a ULCA compliant form and stored.

Another key area where they can support is by aligning Bhashini with the language efforts of individual departments and ministries. These could be through co-branded efforts for crowd sourcing (Bolo India, Padho India etc.) and awareness campaigns through the language missions. They can assist with the following activities:

- Define applications in consultation with the ministries, identification of pain point and accordingly float the challenge round if required

    - Hackathons/Challenges to implement those applications using the Bhashini's resources (such as data/models/compute resources etc.)

- Surfacing of data from different inter-Ministerial departments.
    - Coordinate with other big sources of data like Prasar Bharati, NCERT, Parliament, Judiciary, Digital Libraries etc.
- Coordination with other Missions such as announced by PM-STAIC

6.2. State and Language missions

Language missions shall be established across states to focus on data collection and content creation in specific languages. The language missions shall be region specific, depending upon what language is spoken in the region, and there could be multiple language missions across a single state. It will be the responsibility of the language missions to empanel agencies for data collection, curation and validation activities. All data, whether original or translated, shall be collected using a ULCA compliant process defined by the Data Management Unit.

The language missions shall also be responsible for identification of data sources from state government entities and driving crowd sourcing efforts through standard Bhashini tools. The language missions shall run awareness campaigns for crowd sourcing, which shall be focused on low resource languages. The performance of language missions shall be measured through a public dashboard.

6.3. Academia & Research groups

Academia and research groups shall contribute to the success of Bhashini by carrying out research in the areas of language technologies and translation. The output of all research proposals that are funded by the Government must be open sourced and ULCA compliant. It is further expected that models developed by research must be made available as web (REST) service implementing the ULCA open API. Research proposals should include engineering resources to make this possible.

6.4. Startups

Startups have a key role in Bhashini to develop applications using the open source datasets and models created in Bhashini. Non availability of datasets and models is a major barrier to entry for language technology startups.

Startups will be engaged through challenge rounds and hackathons to build applications based on Bhashini. While datasets and models used are open source, the applications created by them need not be open source. This mission aims to create demand side opportunities for startups that perform well in the hackathons and challenges.

6.4.1. Details of Startup/MSME Engagement

- Start-ups will be supported by providing compute infrastructure to build AI models by leveraging the open source data made available under Bhashini.
- Hackathons/Challenge/Grand Challenge rounds will be held to discover potential of startups
- Workshops will be held to encourage startups to build innovative solutions in language technologies
- Startup/MSME may be empanelled with NHLT for data preparation tasks
- Startup/MSME to be encouraged to register on GEM for providing Localisation services to Govt. Departments

6.4.2. Challenges and Hackathons

Bhashini shall organize challenges and hackathons to develop applications where any Indian start-up (or individual considering to register a start-up) can participate. Applications developed need not be in open source.

- Applications will be measured on following metrics
    - Functionality - Demo
    - Users / Rating
    - Customers
    - Usage of Bhashini’s open source data/models
    - Engineering maturity
    - Other Parameters as defined by Expert Committee

6.4.3. Cloud Credit / API credit distribution

Bhashini shall provide cloud credits for the purpose of model training. This is very important since modern deep learning models require significant computational resources and is a key cost for startups.

- Any startup working in the field of Indic NLP may submit a proposal for Cloud credits for training model and API credits to call Bhashini hosted models
- Proposal will be approved by an expert committee and NHLT will facilitate grant of Compute/API credits
- Startups may request follow-on grants of Cloud/API credit based on a report of the activities and results obtained from the previous grants.
- Grant requests would be evaluated based on the following rubric.
    - Previous experience of the startup in building language AI models and/or applications
    - Any startup having comparable or better than the best benchmarked Bhashini system will be considered favorably.
    - Other Parameters as defined by Expert Committee

6.5. Private companies with large digital reach

Industry has a major role to play in the success of Bhashini in the following ways.

- Industry players with large reach can help in building the data corpus via co-branded crowd sourcing initiatives through their own channels.
- They can help to develop or fund open source software like ‘multilingual content creation platform’ etc. or provide human resources required building the tech backbone of Bhashini.
- They can also contribute by providing compute resources for training and evaluating models.
- They can assist in setting up research collaborations.

The initiatives undertaken by the industry players shall earn them public attribution for contributing to Bhashini. They would also have access to all the open source data and

models developed for Bhashini that they can utilize to build applications and services to solve real world problems.

6.6. Data collection / curation companies

Bhashini shall seek the services of empanelled agencies (and freelancers) for the purpose of data collection, curation and validation. The agencies shall validate the dataset contributions using the process and tools defined by the Data Management Unit. The quality of data collected by each agency shall be measured on an ongoing basis using the maker-checker paradigm. Agencies and freelancers will obtain reputation scores which may be leveraged to obtain additional business, while those with poor reputation scores may be weeded out from the system.

6.7. Publishers

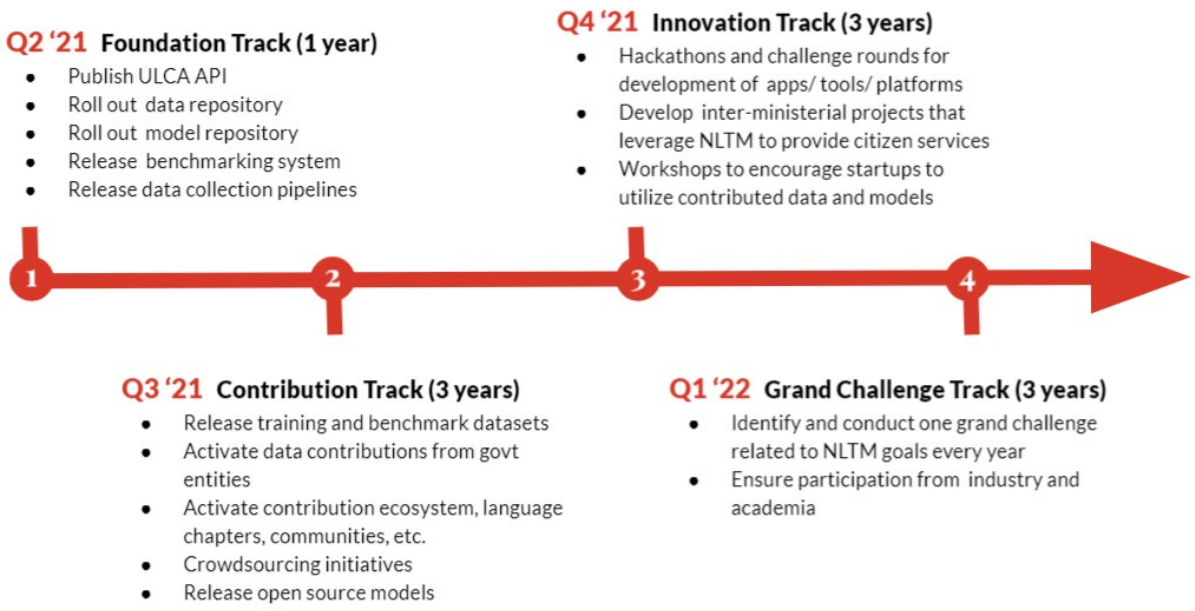
Book publishers are creators, acquirers, custodians, and managers – owners and users – of intellectual property rights. They possess certain rights in the books or any other literary works they produce, and they might also hold other rights on behalf of third parties. Hence, publishers are an important source of useful multilingual data.

In the digital environment, it becomes increasingly important for publishers to protect their copyrights while distributing their works. It is expected that MeitY shall help Bhashini in establishing a working relationship with publishers of multilingual content, in order to utilize sampled data for the purpose of model training while ensuring that the copyrights of the works remain intact. In lieu of their contributions, the publishers shall earn public attribution for contributing to Bhashini. They would also be given access to all the open source data and models developed for Bhashini, to help them build language translation applications for their use.

6.8. Citizens

Language is an emotional issue. And hence, ordinary citizens should be encouraged to contribute to their own language through sachetized crowd sourcing through Bhashini. Bhashini may also help generate job opportunities for people to perform various language related tasks.

7. Action plan



Bhashini is a three year mission which has been segregated across different tracks. During each track, a certain set of activities need to be completed within a specific timeframe for Bhashini to realize its intended objectives from the project.

a) Foundation Track Kickoff (Starting in Q2 2021)

The foundation track shall be active for a period of 12 months and shall aim to set up Bhashini by defining a clear strategy and roadmap for the project. One of the primary activities in this track would be to identify and onboard ecosystem partners. Another important task would be to make the data and contribution foundation ready so that contributors can start providing relevant datasets at the earliest.

The version 1.0 of the data repository and model repository shall be launched at the end of 6 months, along with the foundation layer and the ULCA API, to assist in the data collection and validation process. To catalyze contributions the Bhashini microsite shall also be launched which shall gather ULCA compliant data and also serve as a guide for contributors willing to support Bhashini. The data repository, model repository and the foundation layer shall be upgraded based on our learnings and the version 2.0 of each shall be released by the end of 12 months.

b) Contribution Track Kickoff (Starting in Q3 2021)

The contribution track will be kicked off after the foundation is in place. The contribution track would involve setting up a process to utilize the contributions received from the ecosystem. Bhashini shall ensure that ULCA compliant data submission and tooling is being done by contributors.

The track would be active for three years and will seek to accomplish the following yearly tasks:

i) Year 1

a) Release baseline open source models across 12 major languages.

a) Release training datasets in low resource languages including North Eastern and Tribal languages

b) Create benchmark datasets for all 22 official languages across domains

c) Develop and activate a transparent and continuous benchmarking system

d) Activate data contributions from government entities like Prasar Bharti, NCERT etc and through publically available open source corpora.

e) Sign MoUs with 5 entities each in the public and private sector for data contribution

f) Launch reference crowd sourcing initiatives (Bolo India, Padho India, Likho India, Dekho India)

g) Additional co-branded crowd sourcing efforts with state/language missions and industry partners

ii) Year 2

a) Improve open source baseline for 12 languages

b) Release baseline open source models in 10 additional languages

c) Expand training datasets in low resource languages including North Eastern and Tribal languages and any other languages where sufficient data is not available.

d) Expand benchmark datasets for all 22 major languages across domains to include various NLU tasks such as sentiment analysis, NER, QA, Summarization etc.

e) Sign MoUs with 5 more entities each in the public and private sector for data contribution

f) Sustain state and private-led campaigns for data crowd sourcing, creation and curation. Double the amount of data contributions of crowd sourced compared to Year 1

iii) Year 3

a) Release baseline open source models, create training and benchmark datasets for languages beyond the 22 major Indian languages

b) Sign MoUs with 5 more entities each in the public and private sector for data contribution, specifically for low resource languages

c) Drive campaigns for data crowd sourcing, creation and curation in low resource languages

Mission's Data Collection Goals*	
MT	• 100 million scraped parallel data corpus between Indian Languages • 1,000,000 parallel sentences in 12 languages • 500,000 in 4 languages, 100,000 in remaining 6 languages • 10,000 benchmark quality parallel corpus in each of 22 languages
ASR	• 30,000 hours of unlabeled audio data in Indian Languages • 4000 hours labeled audio data, 2000 hours in English • 1000 hours of labeled audio data in each of 10 languages, 500 hours in 10 languages • 50 hours benchmark quality in each language of 22 languages
TTS	• 40 hours of studio quality audio data in 22 languages
OCR	• 10,000 pages of printed data in each indic script with labeled text • 10,000 images in each Indic script containing scene text
NLU	• 1 million dataset for NER, sentiment analysis, QA for 12 major languages • 100,000 dataset for NER, sentiment analysis, QA for 22 languages
*prior to any data collection exercise (that involves human effort), it shall be confirmed that such data doesn't already exist in the existing datasets.	

Depending upon the progress made with various public and private contributing entities, additional data contribution announcements shall be made during each year.

c) Innovation Track Kickoff (Starting in Q4 2021)

Once the foundation and the contribution tracks are activated, the innovation track shall be kicked off. The innovation track shall also continue for a three year period during which Bhashini shall orchestrate the launch of a market innovation drive for private apps and platforms to co-develop services and solutions. Bhashini shall plan for the following objectives to be achieved during this track:

i) Ensure that at least 5 startups start using the contributed data and models in the first year, for the purpose of development of applications and services

ii) Increase the number of engaged startups to 10 in the second year and subsequently to 20 in the final year

iii) Kickoff co-branded language oriented projects with at least 3 Ministries in the first year

iv) Monitor rollout of projects launched with Ministries in the first year and subsequently kickoff projects with 3 more Ministries in the second year

v) Launch at least one hackathon every quarter

vi) Conduct workshops to encourage startups to utilize contributed data and models.

vii) Undertake activities for inter-ministry alignment, including

a) Discussions on co-funding for language projects in the domains of education, health, finance etc.

b) Driving demand through policy guidelines mandating multi-lingual support in government websites

c) Release of documents and notifications in all Indian languages

d) Discussions with NCERT and state bodies for translation of books and other content

e) Integration with e-Office, reference translation tool, e-Marketplace

f) Activating DIKSHA app for children's voice contribution

g) Kickoff financial inclusion (ATMs, AEPS, BHIM UPI, bank apps) and health inclusion (training, protocols)

d) Grand Challenge Track (Starting in Q1 2022)

Upon the launch of all three tracks vis-a-vis foundation, contribution and innovation, Bhashini will launch grand challenges for promoting innovation in the area of language technology through creation of innovative apps using open source data and models. These events shall be open for both academia and industry wherein they would demonstrate how they can obtain better synergies to fulfill the proposed milestones for the following year, at a better cost.

The objectives for this track shall be as follows:

i) Launch at least one grand challenge per year

ii) Encourage industry and academia partners for participation in grand challenges

iii) Provide funding to the winners of grand challenges for the forthcoming year to meet their proposed milestones.

Delivery of previous year milestones shall be the qualifying criteria for subsequent challenges. This shall enable continued progress on the objectives set forth by Bhashini while ensuring that optimum resources are utilized.

The annual plans for each track shall be reviewed and updated depending upon the progress and Bhashini's learnings from the previous year.

8. Conclusion

Despite having centuries of literary tradition and lakhs of speakers, India's official languages have all been grappling with one common problem – adapting to the demands and the potential of the internet. The issue has proven challenging to solve since it consists of numerous smaller issues like lack of usable data, unavailability of technology resources etc. The importance of solving these issues cannot be understated, given how most Indians simply prefer using their own language in all contexts, including on digital platforms. Their limited support online and on digital platforms has single handedly driven these same Indians to use English instead, even if this means struggling, and ultimately, not being able to make the most of platforms, even for essential services.

Bhashini has developed a roadmap for enabling an ecosystem to make digital content more accessible to the Indian citizens by offering it to them in their native language. Bhashini has introduced certain novel ways in which the ecosystem players can contribute towards the success of this objective. At the same time, Bhashini shall also introduce standards for data collection, curation, benchmarking and adoption. This shall ensure widespread accessibility to data and in turn should result in development of innovative applications and services for the citizens.

With the support of the contributions made by the stakeholders of the ecosystem - namely government, academia, industry and especially the citizens, there will be a multifold increase in the availability of multilingual content. This means that crores of Indians shall be able to use the internet to carry out even the most basic of tasks, without any restrictions. This would enable India to take another leap forward into the digital era.