# National Language Translation Mission (NLTM) Preview

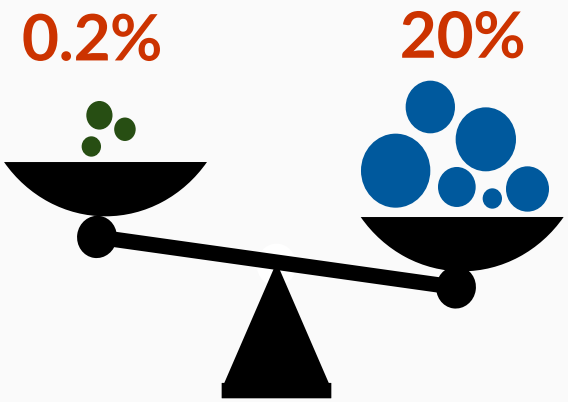
## *Transcending Language Barrier: Digital Inclusion, Digital Empowerment, Atma Nirbhar Bharat*

# Huge gap in content & user base

## Indic Content

**The percentage of websites using Indic language is less than 0.2%**

https://www.forbes.com/sites/niallmccarthy/2018/07/27/how-languages-used-online-compare-to-real-life-infographic/?sh=c9802bb2c7c9



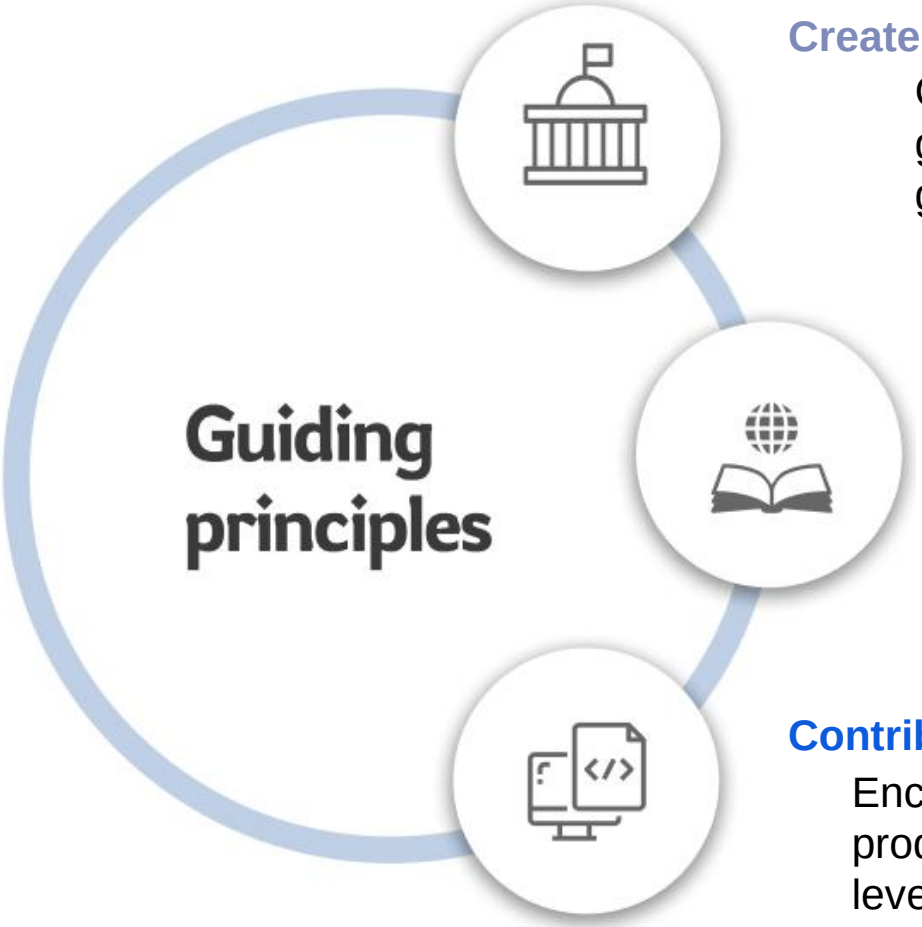

## Huge Population

**With more than 500 million indic language users and close to billion internet users.**

https://assets.kpmg/content/dam/kpmg/in/pdf/2017/04/Indian-languages-Defining-Indias-Internet.pdf

# Mission Statement

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.




## Create an Ecosystem

Create and nurture an ecosystem involving government, industry, academia, research groups, start-ups, and individuals

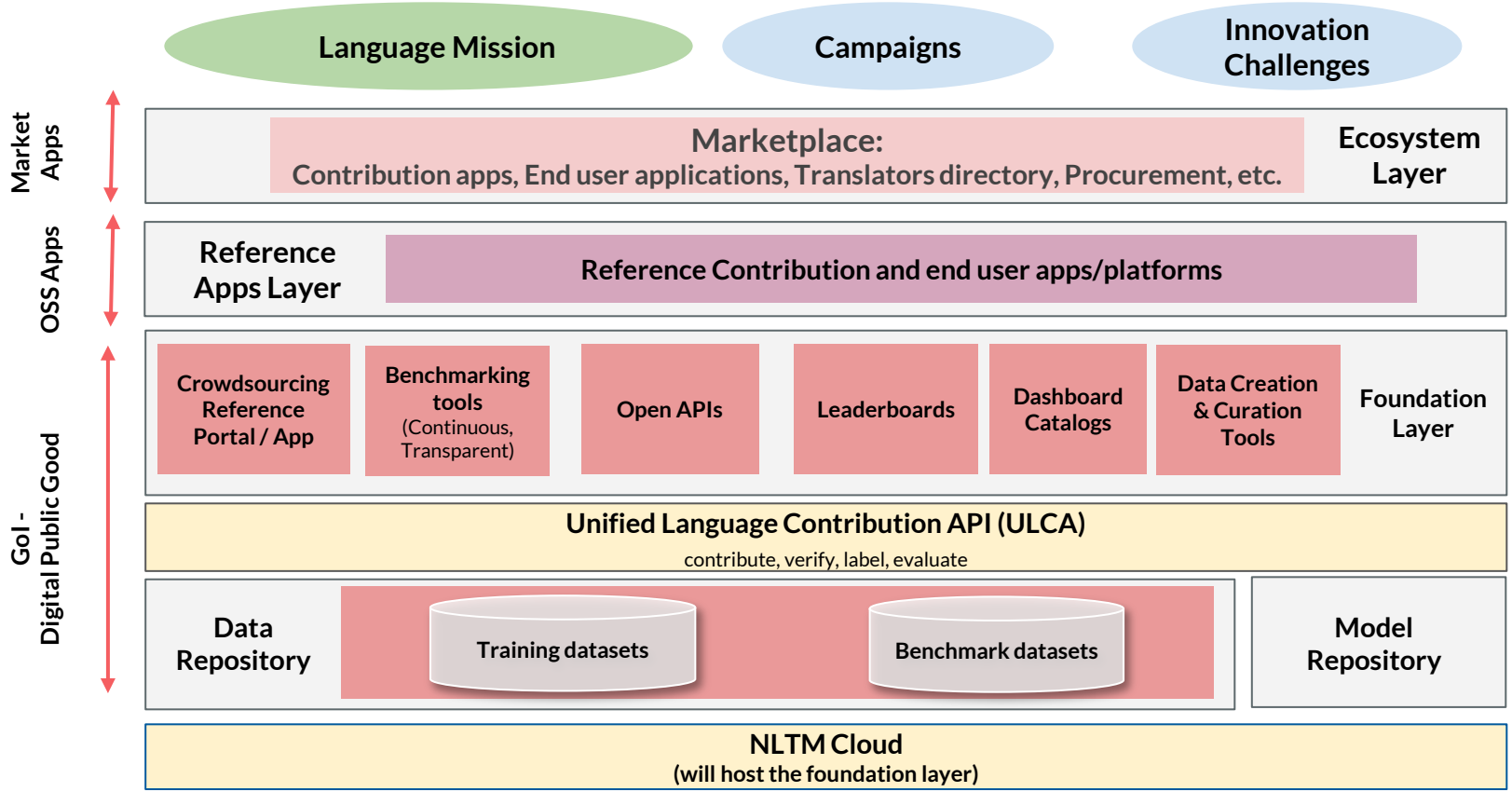
## Open Source Datasets

Create large open source datasets and models by bringing all contributions, both institutional and citizen, into a shared repository that nurtures innovation

## Contribute and Leverage

Encourage the ecosystem to develop innovative products and services in Indian languages by leveraging the open repository of datasets and models.

# NLTM Architecture



# Universal Language Contribution API (ULCA)

ULCA is a standard API and open scalable data platform (supporting various types of datasets) for Indian language datasets and models.

The objective of ULCA is to support research and development of AI tools in Indian languages.

## Open and scalable data platform

- *Parallel text corpus in two or more languages*
- *Monolingual text corpus*
- *Automatic Speech Recognition (ASR) corpus*
- *Text to Speech (TTS) corpus*
- *Optical Character Recognition (OCR) corpus*
- *Natural Language Understanding (NLU) datasets*

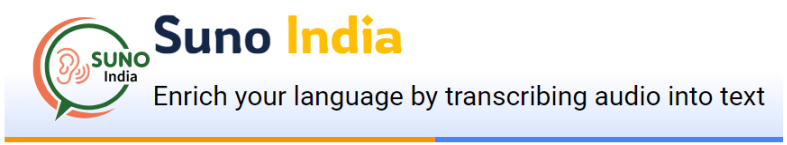
## Indian language models

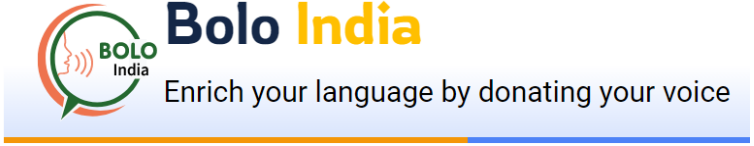
- *Machine Translation (MT)*
- *Automatic Speech Recognition (ASR)*
- *Text to Speech (TTS)*
- *Optical Character Recognition (OCR)*


## Automated Benchmarking

- *Large, diverse and task specific benchmarks*
- *Research community approved metric system*

# Bhasha Daan






**States to mobilize citizens' engagement in crowdsourcing initiatives**

# NLTM Roadmap

## Contribution Track

- *Training and benchmark datasets*
- *Data contributions from government entities,*
- *Language chapters, communities etc.*
- *Crowdsourcing initiatives*
- *Open source language models*

## Grand Challenge Track

- *Conduct one grand challenge related to Bhashini's goals every year*
- *Participation from academia and industry*

## Foundation Track

- *Publish ULCA API*
- *Data repository*
- *Model repository*
- *Benchmarking system*
- *Data collection tools*

## Innovation Track

- *Hackathons and challenge rounds for developing applications*
- *Inter-ministerial projects that leverage Bhashini to provide citizen centric services*
- *Workshops to encourage startups to utilize contributed data and models*

# Mission Ecosystem

### CENTRAL & STATE GOVERNMENTS

Align Bhashini with language specific efforts to attain the mission's objectives

### LANGUAGE MISSIONS

Identify data sources, collect data, create content and plan and execute crowdsourcing initiatives

### ACADEMIA & RESEARCH GROUPS

Engage in research and development activities in language technologies

### STARTUPS

Develop multilingual applications and services

### INDUSTRY

Develop open source software, provide storage and contribute compute for training models

### DATA COLLECTION & CURATION COMPANIES

Collect, validate and curate datasets

### PUBLISHERS

Provide data sources to build datasets and models

### INDIVIDUALS

Contribute to the mission through crowdsourcing initiatives


Digital India
Power To Empower
Thank You
BHASHINI
Bhasha Anek, Bharat Ek