List of ongoing projects pertaining to TDIL Division

S.No.	Document Title (Name of the ongoing Project)	Effective date of the Document/ Start date (Administrative Approval date)	Document Definition	Detailed Guidelines of the Document (Major Deliverable)	Category of the Document
1.	NLTM: BHASHINI	13/10/2021	R&D Project Document	This will enable the wealth of governance-and-policy related knowledge on the internet being made available in major Indian Languages.	Unclas-sified
2.	'OCRs and Applications in Indian Languages' under the Project titled "National Language Translation Mission(NLTM): BHASHINI"	03/02/2022	R&D Project Document	• APIs and technology for public use through a web-based delivery platform as envisaged by the NLTM also co-hosted at IIIT Hyderabad. This will be done for all prominent Indian Languages and all popular character recognition modalities (such as printed, handwriting and scene text). • Data, Annotated data, standards, and public release of the datasets for enabling research and development in the broad space of Indian language OCRs. A portion of the data collection and annotation will be carried out as part of the project. (Additional data will be collected in collaboration with DMU or other agencies suggested by NLTM.) • Manpower trained in this specific domain and catalyzing further technology development outside academia in the future. • 96% accuracy for printed documents across 13 scripts, 94% for handwritten documents across 13 scripts, 92% for scene and video text for 13 scripts.	Unclas-sified
3.	'Indian Language to Indian Language Machine Translation' under the	15/02/2022	R&D Project Document	**Translation Technologies** • English-IL and Indian to Indian Language Machine Translation system (11 language pairs [English<->Hindi, English<->Telugu, Hindi <-> Punjabi, Telugu, Urdu, Gujarati, Kannada, Odia, Kashmiri, Sindhi and Dogri], 22 MT systems)	Unclas-sified

	Project titled "National Language Translation Mission (NLTM) : BHASHINI''			• Domain adapted MT systems for chosen domains [Governance; Educational Content in the fields of Science and Technology (Biology, Chemistry, Physics, Environmental Science, Computer Science Engineering, Electrical Engineering, Mechanical Engineering), Law, Economics, Management; Health Care (Consent Forms and Information Sheets, Awareness and Pharma); Judiciary (Case Files); Agriculture and Food Security] and language pairs; for developing efficient NMT systems approximately 70k parallel corpora for each domain in each language pair is required. **Corpora** • Domain specific parallel corpora for 2 domains and domain dictionaries for chosen language pairs for chosen domains (800k parallel corpora). • Annotated data for chosen domains (Total 180K annotated corpora for chosen languages and domains) **Benchmarks for MT Technologies** • Benchmark standards and guidelines for Indian Languages. • Benchmark data, Methods and Evaluation for MT and MT tools. **Engineering** • API Gateway for Machine Translation engines and utilities. • Productizing IL-IL MT Technologies. **Workshops and Challenge rounds for building ecosystems consisting of language experts and technology developers**	

4.	'Collecting datasets and benchmarks for building Indian Language Technology' under the Project titled "National Language Translation Mission (NLTM) : BHASHINI''	17/02/2022	R&D Project Document	**Task**	**Languages**	**Pretraining tauto)**	**Training (semi-auto)**	**Fine-Tuning**	**Benchmark**	Unclassified
				MT (sentences)	MR LR	10 billion tokens (combined 22 lang.)	1,000,000	100,000 50,000	10,000 10,000	
				ASR (hours)	MR LR	1000 100	1000	100 100	100 100	
				TTS (hours)	MR LR	- -	- -	40 40		
				OCR (documents)	MR LR	- -	100,000 100,000		10,000 10,000	
				OCR (scene)	MR LR	- -	100,000 100,000		10,000 10,000	
				SA (sentences)	MR LR	l0billiontokens (combined 22 lang.)	100,000	10,000	10,000 10,000	
				QA (3uestions)	MR LR	10 billion tokens (combined 22 lang.)	100,000	10,000	10,000 10,000	

				<table><tr><td>NER (sentences)</td><td>MR LR</td><td>10 billion tokens (combined 22 lang.)</td><td>100,000</td><td> 10,000</td><td>10,000 10,000</td></tr></table> Table 1: List of deliverables containing different types of data for each of the five fundamental technology blocks. MR stands for mid-resource languages and includes (Assamese, Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Odia, Punjabi, Tamil, Telugu, Urdu). LR stands for low-resource languages and includes (Bodo, Dogri, Kashmiri, Korikani, Maithili, Manipuri, Nepali, Sanskrit, Santali, Sindhi). Note that only the fine-tuning data and benchmark data will be manually created/curated/verified. All other data (pretraining and training) will be automatically curated from the web. Further all data curated from the web will be released with rules identified by NLTM.	
5.	'Language Communicator Tool for End Users' under the Project titled 'National Language Translation Mission (NLTM) : BHASHINI''	14/02/2022	R&D Project Document	**Language Communicator Tool:** Hindi – Tamil and English i. An API for authoring tool for representing the semantic representation of any language (in this project for Hindi) ii. An aggregate of multilingual generator platform for Tamil and English generators **System Description paper, USR Guidelines**	Unclassified
6.	'English to Indian Language [Hindi, Marathi, Gujarati, Odia, Kannada & Malayalam] and vice versa Machine	24.02.2022	R&D Project Document	• The outcome would be a text-to-text Machine Translation system from English to Hindi, Marathi, Gujarati, Odia, Kannada and Malayalam languages and vice versa • Machine Translation solutions as API/ REST services will be used for further integration to different language-related projects and research works. • Models developed will be available as web REST service implementation ULCA open API	Unclassified

	Translation system' under the project titled 'National Language Translation Mission (NLTM):BHASHINI'				
7.	Discourse Integrated Dravidian Language to Dravidian Language Machine Translation(DL-DiscoMT) under the project titled "National Language Translation Mission (NLTM):BHASHINI'	02.03.2022	R&D Project Document	• A platform for handling Discourse and Conversation • A text to text Machine Translation system from Hindi to Tamil, Tamil to Hindi, Kannada, Malayalam and Telugu Bi-direction systems. Incorporating discourse information in NMT and Sampark. • Leaderboard platform for Evaluation • Machine Translation solution as API/services which can be used for integrating with SSMT systems and by end users.	Unclassified
8.	'Speech technologies in Indian languages' under the Project titled "National Language Translation Mission (NLTM): BHASHINI'	18.02.2022	R&D Project Document	Setting up standards for data collection, curation, archival, using best practices and benchmarks adapted for the Indian language. 1) ASR • ASR systems in Indian English, Tamil Hindi, Telugu, Bengali, Gujarati, Marathi, Assamese, Kannada, Malayalam, Odia, Punjabi (Tonal language), Bodo (Low Resource language) and Manipuri • Total ASR Corpus size for above languages: 30000 hours • 8,000 hours of NPTEL Indian English Technical data curation 2) TTS • TTS systems in Hindi, Tamil, Indian English, Marathi, Bengali,	Unclassified

				Malayalam, Telugu, Assamese, Kannada, Gujarati, Odia, Rajasthani, Bodo, Manipuri, Urdu, Punjabi, Kashmiri, Konkani. • Total TTS Corpus size for above languages: 1360 hours. • Develop voice search and voice assistant in Indian English and Hindi.	
9.	'Speech Datasets and Models for Tibeto-Burman Languages(SpeeD-TB)' under the Project titled "National Language Translation Mission (NLTM): BHASHINI'	22.02.2022	R&D Project Document	The aims and objectives and complete deliverables of the project are as listed below - • To build a transcribed speech dataset of approximately 200 hours each in 6 Tibeto-Burman languages - Bodo (mainly spoken in Assam), Meetei (mainly spoken in Manipur), Chokri (mainly spoken in Nagaland), Kok Borok (mainly spoken in Tripura), Nyishi (mainly spoken in Arunachal Pradesh) and Toto (mainly spoken in West Bengal) • To develop a phone set for each of the languages under study. • To build a language model for the languages under consideration here. • To build a baseline ASR system for each of the above languages. • To make the dataset and pre- trained and fine-tuned models publicly available through Bhashini/ ULCA and also other platforms and sources including GitHub and other appropriate repositories and server under CC-By 4.0 license (for dataset) and AGPL v3 (for the model).	Unclassified
10.	An Interpretable Unified Framework for Text-to-Text Translation among Indian Languages using Sanskrit-	25.03.2022	R&D Project Document	**T2T translation systems** (Language pairs corresponding to Sanskrit, Hindi, Kannada): The interlingua-based translation models for improved interpretability and faithfulness: API and Web-interface. The models are expected to be more accurate than the available open-source models at ULCA platform, or the Indic-trans platform. **Linguistically rich annotated data for the 3 languages**: 40k sentences per language, which will be annotated as per	Unclassified

	based Interlingua Representation' under the Project titled "National Language Translation Mission (NLTM) : BHASHINI'			the interlingua annotation scheme with the help of the available morphology tools. The data will be released under CC-BY 4.0 license and can be used for any purpose by all on ULCA.	

11.	'VIDYAAPATI Bidiractional Machine Translation involving Bengali, Konkani, Maithili, Marathi and Hindi' under the Project titled "National Language Translation Mission (NLTM) : BHASHINI'	30.03.2022	R&D Project Document	• Bidirectional MT system: Hindi - Bengali, Konkani, Maithili, Marathi.
				• Mobile App, Web-service, and APIs of the MT systems.
				• Linguistic Resources:
				  • Domain-wise size:
				  • Governance and Policy including Judiciary: 50%
				  • Education: 30%
				  • Rest (Science and Technology, Healthcare, Agriculture, Climate, Tourism, etc.): 20%
				  • Parallel corpora for each language pair (Hindi - X, where X is one of Bengali, Konkani, Maithili, Marathi) of approximate size 25K parallel sentences will be created. (This is as per MEITY's instruction; 10% of the data will be created by the consortium and 90% by the DMU)
				  • MW, NE, and POS tagged corpus of approximate size 25K sentences of each language among Bengali, Konkani, Maithili, and Marathi.
				Open-source: code, data, and models will be available to the community for development and utilization. The source code will be released under the license AGPLv3 or Mozilla-v2. The Data will be released under CC-BY 4.0 license. The data and the models will be uploaded to the ULCA.
					Unclassified

				Evaluation metrics and framework. Deployment strategy in language technology.	

12.	'ISHAAN: A system for Bidirectional Machine Translation between
	1) English and Assamese ,Bodo, Manipuri, Nepali
	2) Manipuri and Hindi
	3) Assamese and Bodo'
	under the Project titled "National Language Translation Mission (NLTM) : BHASHINI'
		30.03.2022	R&D Project Document	• Bidirectional MT systems:
				  • English - Assamese, Bodo, Manipuri, Nepali
				  • Hindi - Manipuri
				  • Assamese – Bodo
				• Mobile App, Web-service, and APIs of the MT systems.
				• Linguistic Resources:
				  • Parallel corpora for each language pair (English - 4 North-East Indian Languages) (Approximately 25K parallel sentences for English-X, Hindi- Manipuri, and Assamese- Bodo, where X is one of Assamese, Bodo, Manipuri, Nepali)
				  • Domain-wise size:
				  • Governance and Policy including Judiciary: 50%
				  • Education: 30%
				  • Rest (Science & Technology, Healthcare, Agriculture, Climate, Tourism, etc.): 20%
				  • Multiwords (MW), Named Entity (NE) and Part-of-speech (POS) tagged corpus of size approximately 25K sentences for each North-East Indian language.
				Open-source: code, data, and models will be available to the community for development and utilization. The source code will be released under the license AGPLv3 or Mozilla-v2. The Data will be released under CC-BY 4.0 license. The data and the models will be uploaded to the ULCA.
				Evaluation metrics and framework.
				Deployment strategy in language technology.
					Unclassified

13.	'National Hub for Language Technology'	23.09.2022	R&D Project Document	• To develop, enhance & manage the BHASHINI (BHASHaINterface for India) Platform, leveraging open-source components, Government built solutions,	Unclassified

	under the Project titled "National Language Translation Mission (NLTM) : BHASHINI'			IndEA components, etc. • To provide support related to engineering related aspects to different NLTM units viz. R&D institutions, Data Management Unit (DMU), Ecosystem Engagement Unit (EEU) etc. for focused and time-framed development.	
14.	'Sanskrit Knowledge Accessor' under the Project titled "National Language Translation Mission (NLTM) : BHASHINI'	21.03.2024	R&D Project Document	• Three Machine Translation cum Accessors from Sanskrit-Hindi, Sanskrit-English and Sanskrit-Tamil would be developed and deployed on a cloud server. • An Ayurveda reading aid for Vaidyas, and • Darshana or classics reading aid for popularization among reading public such as Mahābhārata.	Unclassified
15.	'Establishment of BHASHINI Ecosystem Engagement Unit (EEU) for Startups and Industries' under the Project titled "National Language Translation Mission (NLTM) : BHASHINI'	07.12.2023	R&D Project Document	• To Develop different reference applications in language technology based on Bhashini tools in the area of Translation, transliteration, NER, ASR, TTS, OCR and language detection etc. (by startups/MSME & Industries) • To develop new AI models in Indian language combinations that are not available in the Bhashini pool (by startups/MSME & Industries) • To develop large language models/ new technology-based models in Indian language combinations and trained through Bhashini database for all domains and their specific sub domain, so that model can be usable in any area specific application. (by startups/MSME & Industries) • To trained Bhashini model or other open source model through Bhashini database for different domains and their specific sub domain, so that model can be usable in any area specific application. (by startups/MSME & Industries). • Bhashini will provide curated data, computing and Bhashini platform support to startups.	Unclassified

16.	Establishment of BHASHINI 'Ecosystem Engagement Unit (EEU) for State and Central Government' under the Project titled "National Language Translation Mission (NLTM) : BHASHINI'	13.03.2024	R&D Project Document	• To establish a Bhashini Ecosystem Engagement Unit to coordinate with the State Language Missions(SLMs) and the Central Government working together to develop and deploy innovative products and services in Indian Languages. • To Develop different reference applications in coordination with State and Central Government capturing their requirements for SLMs in language technology based on Bhashini tools in the area of Translation, transliteration, NER, ASR, TTS, OCR and language detection etc.	Unclassified