

**Pre-Standardization Studies for Indian Language  
Resources**  
**Industry Connections Activity Initiation Document (ICAID)**  
**Version: 1.0, 4 September 2019**

IC19-007-01 Approved by the IEEE-SASB 5 September 2019

### Instructions

- Instructions on how to fill out this form are shown in red. It is recommended to leave the instructions in the final document and simply add the requested information where indicated.
- **Shaded Text** indicates a placeholder that should be replaced with information specific to this ICAID, and the shading removed.
- Completed forms, in Word format, or any questions should be sent to the IEEE Standards Association (IEEE-SA) Industry Connections Committee (ICCom) Administrator at the following address: [industryconnections@ieee.org](mailto:industryconnections@ieee.org).
- The version number above, along with the date, may be used by the submitter to distinguish successive updates of this document. A separate, unique Industry Connections (IC) Activity Number will be assigned when the document is submitted to the ICCom Administrator.

### 1. Contact

Provide the name and contact information of the primary contact person for this IC activity. Affiliation is any entity that provides the person financial or other substantive support, for which the person may feel an obligation. If necessary, a second/alternate contact person's information may also be provided.

**Name:** Kalika Bali

**Email Address:** kalikab@microsoft.com

**Employer:** Microsoft Research

**Affiliation:**

IEEE collects personal data on this form, which is made publicly available, to allow communication by materially interested parties and with Sponsors and Activity officers who are responsible for IEEE work items.

### 2. Participation and Voting Model

Specify whether this activity will be entity-based (participants are entities, which may have multiple representatives, one-entity-one-vote), or individual-based (participants represent themselves, one-person-one-vote).

Specify: "Entity-Based"

### **3. Purpose**

#### **3.1. Motivation and Goal**

Briefly explain the context and motivation for starting this IC activity, and the overall purpose or goal to be accomplished.

Identification of use cases that could lead to proposals for standards to govern Language Resources for Indian languages is the motivation of this effort. Language Resources include Speech, Language data and descriptions which are made available in a machine readable form and used for developing, evaluating and improving algorithms in the area of natural language and Speech processing. Such standards can also be used for language studies, localization of software, language studies, electronic publishing and any purpose for researchers, subject area specialists etc. Examples of Language Resources are spoken and written corpora, computational lexica, terminology databases etc. The work will also include any tools that are used for the above stated purposes.

#### **3.2. Related Work**

Provide a brief comparison of this activity to existing, related efforts or standards of which you are aware (industry associations, consortia, standardization activities, etc.).

The work here is very similar to the effort undertaken by European Language Resource Association ([ELRA/ELDA](#)). In India, TDIL has also undertaken standardization efforts for Indian languages which should be included and extended wherever required as part of this effort.

#### **3.3. Previously Published Material**

Provide a list of any known previously published material intended for inclusion in the proposed deliverables of this activity.

List the previously published material, if any. Refer to TDIL [standardization](#) page.

#### **3.4. Potential Markets Served**

Indicate the main beneficiaries of this work, and what the potential impact might be.

India is the primary market for the standards defined as part of this effort. Further, the standards can be used in any geo where any of the Indian languages are used or if the respective language(s) is derived from the same root as Indian languages.

#### **3.5. How will the activity benefit the IEEE?**

Language enabled products are gaining wide traction in every single sector of the economy. It is acknowledged that the next Billion internet users in India will come as a result of localization of contents in Indian languages. Translation, Text recognition, Context aware automation, extraction etc. are the key use cases driving the deployment. The biggest challenge and hence an area of active innovation for Indian language is that most of these languages are under resourced.

All this combined opens up a big opportunity for standardization and fits very nicely with why IEEE drives standardization activities. Standardization here will spur innovation, fuel new growth of technology consumers, expand customer choice, reduce cost structures and support interoperability between multiple languages among other benefits.

#### **4. Estimated Timeframe**

Indicate approximately how long you expect this activity to operate to achieve its proposed results (e.g., time to completion of all deliverables).

**Expected Completion Date:** 12/2021

IC activities are chartered for two years at a time. Activities are eligible for extension upon request and review by ICCom and the IEEE-SA Standards Board. Should an extension be required, please notify the ICCom Administrator prior to the two-year mark.

#### **5. Proposed Deliverables**

Outline the anticipated deliverables and output from this IC activity, such as documents (e.g., white papers, reports), proposals for standards, conferences and workshops, databases, computer code, etc., and indicate the expected timeframe for each.

Specify the deliverables for this IC activity, please be specific. Expected deliverables include:

1. List of Language Resources required for the officially recognized Indian languages
2. Identification of Standards for each of the identified Language Resources
3. Database with sample collateral for each of the listed Language Resources across official Indian languages
4. Computer code for a sample list of common use cases using the Language Resources made available in database
5. Workshops, conferences etc. for discussions, deliberations, paper submission contests etc. as appropriate

#### **6. Funding Requirements**

Outline any contracted services or other expenses that are currently anticipated, beyond the basic support services provided to all IC activities. Indicate how those funds are expected to be obtained (e.g., through participant fees, sponsorships, government or other grants, etc.). Activities needing substantial funding may require additional reviews and approvals beyond ICCom.

No additional funding requests are anticipated for services beyond the standard services provided for IC programs. Activity members will provide any needed support for hosted meetings, marketing activities that exceed basic IC support.

#### **7. Management and Procedures**

### **7.1. IEEE Sponsoring Committee**

Indicate whether an IEEE sponsoring committee of some form (e.g., an IEEE Standards Sponsor) has agreed to oversee this activity and its procedures.

**Has an IEEE sponsoring committee agreed to oversee this activity?:** NO

If yes, indicate the sponsoring committee's name and its chair's contact information.

**Sponsoring Committee Name:** Committee Name

**Chair's Name:** Full Name

**Chair's Email Address:** who@where

Additional sponsoring committee information, if any.

IEEE collects personal data on this form, which is made publicly available, to allow communication by materially interested parties and with Sponsors and Activity officers who are responsible for IEEE work items.

### **7.2. Activity Management**

If no IEEE sponsoring committee has been identified in 7.1 above, indicate how this activity will manage itself on a day-to-day basis (e.g., executive committee, officers, etc).

The activity will be managed by an executive committee as defined in the activity's policies and procedures.

### **7.3. Procedures**

Indicate what documented procedures will be used to guide the operations of this activity; either (a) modified baseline *Industry Connections Activity Policies and Procedures*, (b) Sponsor policies and procedures accepted by the IEEE-SA Standards Board, or (c) Working Group policies and procedures accepted by the Working Group's Sponsor. If option (a) is chosen, then ICCom review and approval of the P&P is required. If option (b) or (c) is chosen, then ICCom approval of the use of the P&P is required.

Will use the baseline Industry Connections Activity Policies and Procedures.

## **8. Participants**

### **8.1. Stakeholder Communities**

Indicate the stakeholder communities (the types of companies or other entities, or the different groups of individuals) that are expected to be interested in this IC activity, and will be invited to participate.

Participation is grouped under the following categories: Government, Academia and Industry. The list below limits itself to the names of the institutions and does not (yet) identify individuals within them. Note that this is an initial list and not intended to be read as the complete list.

**Government:**

1. TDIL
2. FICCI
3. CDAC

**Academia:**

1. IIT Mumbai/Delhi/Chennai/Patna
2. IIIT Hyderabad/Bangalore
3. JNU Delhi

**Industry**

1. Microsoft Research
2. Google India
3. Amazon India
4. TCS Research
5. Intel Technology India Pvt. Ltd
6. Reliance Jio
7. Flipkart
8. NVidia

**8.2. Expected Number of Participants**

Indicate the approximate number of entities (if entity-based) or individuals (if individual-based) expected to be actively involved in this activity.

Number of entities or number of individuals.

**8.3. Initial Participants**

Provide a number of the entities or individuals that will be participating from the outset. It is recommended there be at least three initial participants for an entity-based activity, or five initial participants (each with a different affiliation) for an individual-based activity.

Use the following table for an entity-based activity:

<b>Entity</b>	<b>Primary Contact</b>	<b>Additional Representatives</b>
Entity Name	Contact Name	Name
Microsoft Research	Kalika Bali	Monojit Chaudhury
FICCI	Sarika Gulyani	
Intel Technology India Pvt. Ltd	Raghavendra Bhat	