Sixth Meeting – minutes

Minutes of the Sixth Project Advisory Committee Meeting of the Linguistic Data Consortium for Indian Languages (LDC-IL) held on March 16, 2018 at CIIL, Mysore

MEMBERS PRESENT		
1.	Prof. D.G. Rao Director Central Institute of Indian Languages Manasagangotri, Hunsur Road, Mysore - 570 006	Chairperson
2.	Prof. D.G. Rao	Chairperson
3.	Dr. Narayan Kumar Choudhary	Officer i/c, LDC-IL
4.	Prof. Dipti Misra Sharma	IIIT, Hyderabad
5.	Prof. Swaran Lata	Head, TDIL, MCIT, New Delhi
SPECIAL INSTITUTIONAL INVITEES		
6.	Dr. Anil Kumar Singh	IIT - BHU, Varanasi
7.	Dr. Ajai Kumar	C-DAC, Pune
8.	Dr. S.R. Savithri	AIISH, Mysore
SPECIAL INDIVIDUAL INVITEES		
9.	Prof. Aadil Amin Kak	University of Kashmir, Srinagar
10.	Dr. Niladri Sekhar Dash	ISI, Kolkata
11.	Dr. Kamal Kumar Choudhary	IIT Ropar, Punjab
SPECIAL INDUSTRY INVITEES		
12.	Mr. Krishna Doss M	MICROSOFT
13.	Mr. Daan Van Esch	Google
14.	Ms. Amrita Kamat	Google
15.	Mr. Manish Chapla	Keypoint Technologies, Hyderabad
16.	Mr. Vivekananda Pani	Reverie Technologies, Bangalore
17.	Mr. Bhupen Chauhan	Reverie Technologies, Bangalore
18.	Ms. Juby Jose	Intel Corporation, Bangalore
19.	Ms. Dona Sihi	FICCI-ILIA
20.	Ms. Satabdi Sengupta	FICCI-ILIA

Some members attended the meeting via Video Conferencing. Their names are as follows:

1.	Prof. Pushpak Bhattacharya Indian Institute of Technology, Patna
2.	Prof. Girish Nath Jha, JNU, New Delhi

Welcome

Dr. L. Ramamoorthy, Head, Centre for Corpus Linguistics, CIIL welcomed the members of the 6th Project Advisory Committee Meeting.

Prof. D. G. Rao, Director, CIIL addressed the meeting and set up the motion to give a brief of CIIL and LDCIL. Given that the PAC meeting is taking place almost after 6 years, most of the members are new and they were apprised of this initiative running at the Institute for a long time.

Agenda Items and detailed brief about the works of LDCIL in the past ten years were presented by Dr. Narayan Choudhary, Officer in-charge, LDCIL.

Thereafter, agenda items were taken one by one.

Costing Plan for the Data Sets:

The costing plan has also been an issue being considered since the last PAC meetings and through the sub-committees. As costing plan has to be universal and the data set should be made available to both the research and industry, a universal pricing plan should be made that is justifiable and attractive to the research and development community.

It is suggested that the cost of a language resource should be made as per the current cost of creating such a resource. It may happen that the cost of creating a resource may be higher for the agency that has developed it due to various reasons. Those reasons need not be applicable to the prospective users of the data sets. Therefore, a reasonable pricing plan has to be chalked out for each of the data sets.

The matter has been a long standing issue with the TDIL, Meity as well which has been struggling to make its data available for commercial use.

To resolve this issue, it was suggested by TDIL and CIIL that a policy document be prepared such that it contains a universal pricing plan indicating all the factors affecting the cost of a language resource development for language technology.

Dr. Narayan Choudhary, OiC, LDCIL was entrusted by CIIL to develop such a document with feedback from different stakeholders and based on his own experiences of working on these types resources. The cost analysis document was shared with the Meity as well as all the members of this PAC, with necessary feedback.

The PAC endorses the document to be used as a guiding document to arrive at the cost of the various language resources subject to the following changes:

1. The DTP rates, Proofing rates and translation rates for difficult domains (e.g. scientific and technical nature) may be revised as the rates are rather old (2012). This may be ratified in the next meeting..
2. Cost of segmentation/tokenization for inflecting and agglutinating languages may be factored into the formula.
3. Cost of mediator/interpreter for the field based works (e.g. speech data collection) may also be included.
4. Expenses on Travelling and Accommodation may also be factored in.
5. The section on dependency/syntactic parsing may be removed and finalized after discussions with Prof. Dipti Mishra Sharma, IIIT-H and Dr. Niladri Shekhar Dash, ISI, Kolkata.
6. Other sections on WordNet etc. will also be added.

This document may be published as an e-book and made available on LDCIL and CIIL websites. Necessary updates on this document from time to time may be done by CIIL based on the new feedback coming from different stakeholders.

The actual cost finalization for each of the data set based on the formula would be done by the implementing/developing agencies based on the parameters as set out in the formulae given. The parameter where a price is fixed in the formulae is subject to change based on the extant conditions as applicable to the agencies/organizations developing the language resource.

Price Tiers:

The issue of making the data sets available to the research and industry for a price or without a price was discussed at length in the meeting. The following pointers were taken into consideration:

1. LDCIL was established with a goal become self-sustained after six years of operation and funding from the Central Government.
2. LDCIL is expected to generate revenue out of selling the language resources developed by it.
3. It was observed in the previous PAC meeting that the goal of self-sustenance is highly impossible for LDCIL as the funds generated out of it would not be enough to make it possible.
4. However, LDCIL must strive to generate funds by engaging with the commercial entities.
5. To promote Indian languages and open source knowledge, the data sets may made free for academic research and not-for-profit organizations. This will help faster technology development for the Indian languages where required resources scarce and language technology developers/enthusiasts may benefit out of it.
6. Revenue should be generated from prospective commercial entities interested in using the data sets for such purposes.
7. Two pricing tiers may be applied for commercial entities, one for Macro, Small and Medium level Enterprises (MSMEs) and one for Large Scale Enterprises (including MNCs) as defined by the GoI
8. In future, collaborative efforts for development of data may be explored.

Based on the pointers given above, the committee felt that it requires more discussion and suggested that the following proposals may be discussed among the committee members over the next few months over email exchanges and finalized in the next PAC meeting to be held within a period of three months.

- Whether Language Resources may be set as free for academic research and not-for-profit organizations.
- Whether Cost of a corpus may be set at 80% of the overall cost of a data set for Large Scale Enterprices and MNCs.
- Whether Cost of corpus may be set at 60% of the overall cost of a data set for MSMEs.
- Whether special pricing tier should be made for startup companies.
- How researchers from the international community can be engaged/benefited.

Licensing Issues:

It is agreed that the licensing documents may be prepared based on the model of licensing documents used by the TDIL programme of Meity. The following licensing documents are to be

prepared and discussed over next few months over email exchanges and finalized in the next PAC meeting.

- Proforma for Terms & Conditions.
- Privacy Policy of LDCIL Website (to be updated).
- Undertaking for Academic Research and Not-for-Profit Use.
- Certification for Research and Not-for-Profit Use.
- Certification for MSMEs.
- Certification for LSI and MNCs (may be skipped if declared by the applicants).

In addition to the above, the LDCIL website will host a catalogue of its language resources offered on a sub-domain at http://data.ldcil.org. This sub-domain will be modeled on the catalogue page of LDC, Upenn or ELRA and contain all the peripheral information about the data set, including prices.

Necessary upgrades on the LDCIL website will be carried out.

Copyright Issues:

It was found that the LDCIL has a huge resource for the Indian languages which has not yet been seen by the research and development community of language technology. The main reason for this delay has been the copyright permission which was not sought from the respective copyright holders.

The issues of copyright and licensing were discussed in the last PAC meetings as well as in the licensing and pricing sub-committees. Following the actions suggested in those meetings, advice from the Ministry was sought as to copyright issues whereupon the LDCIL was advised that copyright permission is a must before the data is released to the public for any use, be it research or commercial.

This is to note that the task of procuring the permissions to use the extracts of the books/titles is a cumbersome task as this copyright information were included earlier when the data was selected and made part of the text corpora. Thus, it is an additional task to identify the copyright holders of those titles and then find their contact details before approaching them for the necessary extract permissions.

Anyway, LDCIL has already started this process. There are more than 90,000 titles for which permissions have to be sought. Once copyright permissions are sought for a data set, it can be released.

With the current efforts, it is found that the Odia text corpus already has collected the necessary permissions for more than 80% of its titles and is ready to be released. Therefore, if decided, Odia text corpus may be released as soon as possible.

It is noted that the Speech corpora do not have any copyright issues as the data is solely owned by CIIL. Therefore, speech data sets may also be made ready for release and be released as soon as possible.

Data Publishing Policy:

- Each data set will have a unique identifier as well as an ISBN number (issued by the CIIL) for all of resources of LDCIL and TDIL
- Data Sets to contain the documentation, file structure, necessary licenses and other peripheral information along with the main data set itself as well as the metadata.
- LDCIL portal to receive requests for data sets and deliver it there following necessary terms and conditions.
- The data distribution portal may be set functional ASAP and placed before the PAC again before data sets are released..
- Due to credits to the people/resource person contributing to it may be given (includes the respective resource personnel, OiC and Head of LDCIL for its own data set or other respective entities for data sets prepared outside of LDCIL).

The meeting ended with the vote of thanks to the chairperson.

Sd/-

(Prof. D.G. Rao)

Chairperson, LDC-IL & Director, CIIL