SO MANY LANGUAGES: CHALLENGES AND OPPORTUNITIES FOR THE WIKIMEDIA MOVEMENT IN INDIA

'Diversity Conference' November 9, 2013

https://meta.wikimedia.org/wiki/Wikimedia_Diversity_Conference

Berlin, Germany

T. Vishnu Vardhan

Prg. Director Access to Knowledge, Centre for Internet & Society

✉ vishnu@cis-india.org   vishnuvardhan_t

Looking at the Language Diversity in India

Challenges and Opportunities

WikiSpeech – Project idea ☺

✉ vishnu@cis-india.org   vishnuvardhan_t

Half of all the moderns languages of the world my vanish by 2100 – UNESCO Report, 2010*

In history, very large languages also go down sometimes. Latin is one example. The (ancient) Greek language is another, Sanskrit is the third one. A language does not have to be small in order to face extinction. That is the nature of language – Ganesh Devy**

✉ vishnu@cis-india.org   vishnuvardhan_t

Language Diversity in India – A historical Glimpse

➤ Linguistic Survey of India – George A. Grierson (1894-1928)

➤ Context: After the Great Uprising of 1857, the reestablishment and deeper penetration of the colonial state in India resulted in the systematic collection and codification of knowledge about things and people Indian – Prof. Shahid Amin (2008) *

➤ 179 languages and 544 dialects in the Indian Empire.

➤ 1961 Census reports – 1652 mother tongues**

* http://dsal.uchicago.edu/lsi/content/Introduction
** http://www.languageinindia.com/aug2002/indianmothertongues1961aug2002.html

✉ vishnu@cis-india.org   vishnuvardhan_t

Language Diversity in India

➢ 122 languages spoken by > 10,000 - Census of India 2011

➢ Of these the Constitution of India recognizes 22 languages

➢ About 87 languages used in print media.

➢ 71 languages used on radio

➢ Administration in 15 languages

➢ Media of instruction 47 languages*

* Data sources

 Census of India Report , 2011

Rao, S. 2008. India's language debates and education of linguistic minorities. Economic and Political Weekly, 6 September 2008.

Meganathan, Ramanujam. Language policy in education and the role of English in India: From library language to language of empowerment.
https://www.teachingenglish.org.uk/sites/teacheng/files/Z413%20EDB%20Section04_0.pdf

✉ vishnu@cis-india.org   vishnuvardhan_t

People's Linguistic Survey of India

- ➤ 780 languages
- ➤ 66 scripts (e.g. Konkani)
- ➤ Arunachal Pradesh is the richest with 90 languages
- ➤ Survey by more than 3,000 volunteers over 4 years
- ➤ 35,000-page survey in 50 volumes
- ➤ **220 languages died in the last 50 years**
- ➤ **150 could die soon**

* Sources
http://blogs.reuters.com/india/tag/peoples-linguistic-survey-of-india/
http://www.thehindu.com/news/national/language-survey-reveals-diversity/article4938865.ece

✉ vishnu@cis-india.org   vishnuvardhan_t

Indian Languages on Wikimedia

➢ Wikipedia project in 22 languages

➢ More than 20 Indian language Wikipedias in incubation

➢ More details

http://wiki.wikimedia.in/List_of_Indian_language_wiki_projects

✉ vishnu@cis-india.org

vishnuvardhan_t

✉ vishnu@cis-india.org   🐦 vishnuvardhan_t

Wikipedia

286 Language Wikipedias
334 Language Wikipedias in incubation
Total Articles – 2,60,00,000
Articles in English Wikipedia – 42,00,000
16,43,063 Editors/Contributors
73,049 Active Editors/Contributors
36,50,00,0000 Page Views per month

THE CENTRE FOR internet & society

Access To Knowledge

Top 5 Indian Wikipedias have 2,56,160 Articles

Total Indian Contributors – 5514

Very Active Editors – 467

2,75,79,239 page views per month

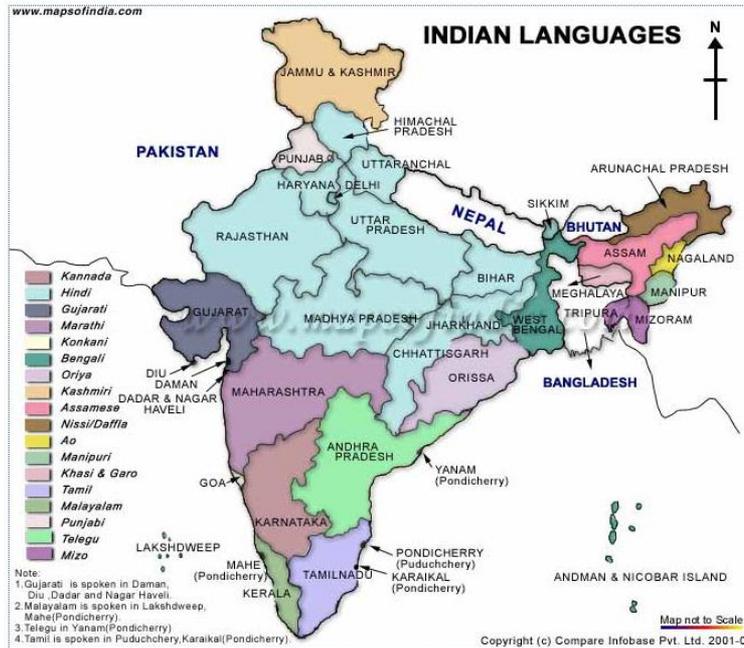## Indian Language Wikipedia - Factsheet
Indian Language Wikipedia Statistics - as on March 31, 2013

	Hindi	Bengali	Punjabi	Marathi	Telugu	Tamil	Urdu	Kannada	Gujarati	Malayalam	Oriya	Assamese	Kashmiri	Sanskrit
Page Views per Month	5,926,278	3,028,411	344,503	2,913,713	1,456,140	3,434,075	1,291,857	891,997	666,853	2,294,413	292,777	349,024	54,421	383,750
Article Count	97,151	25,725	5,690	39,313	52,522	52,431	23,670	14,255	22,611	30,250	3,690	1,990	113	10,272
New Articles per Day	7	7	7	9	8	21	13	11	2	23	6	2	0	1
Edits per Month	33,874	29,531	5,720	8,758	16,239	47,551	25,450	10,113	12,632	29,077	4,142	2,485	132	6,306
Active Editors	44	61	10	29	36	86	26	25	20	124	13	14	0	13
Very Active Editors	9	8	3	8	12	20	5	6	2	26	4	2	0	0
New Editors	14	17	3	8	6	20	4	12	3	19	1	0	0	5
Speakers	550,000,000	230,000,000	104,000,000	90,000,000	80,000,000	66,000,000	60,000,000	47,000,000	46,000,000	37,000,000	31,000,000	13,000,000	4,600,000	50,000
Editors per Million Speakers	0.1	0.3	0.1	0.3	0.5	1	0.4	0.5	0.4	3	0.4	1	0	0

Challenges – Knowledge Production in Indian Languages

- Orality of the language (e.g. Sea in Coastal fishermen's language)
- Knowledge production in Indian Languages (e.g. Urdu)
  - a) individual efforts mostly on creative side
  - b) state funded knowledge production (e.g. Konkani)
- The aspiration of English education
- Relative lack of research/academic standards, which is transferred on to Indic Wikipedias
- Translation challenges (EN-Indic-EN; Indic-EN-Indic; Indic-Indic)
- Regional identity to language (see map)

(vs)

In all the states and union territories, invariably, there are at least 10 languages or more, but in urban spaces, like Delhi or Bombay, or Hyderabad or Bangalore, nearly 300 language communities inhabit that space in a very substantial number for each community. In the northeast, there are more than 250 languages. – Ganesh Devy, PLSI

Image source: www.management-class.co.uk

✉ vishnu@cis-india.org   vishnuvardhan_t

Challenges for Wikimedia growth in Indian Languages – Efforts of CIS-A2K

- Technical infrastructure for Indian Languages
  - Inadequate support to Indian scripts on digital environment [Lobbying with tech players and Govt. policy change]
  - Typing a major challenge (too many letters too less keys) [video tutorials; Indic typing at school level]
  - OCR – complexity of Indian Language scripts [Lobbying with C-DAC. Exploring support to tesseract based development]
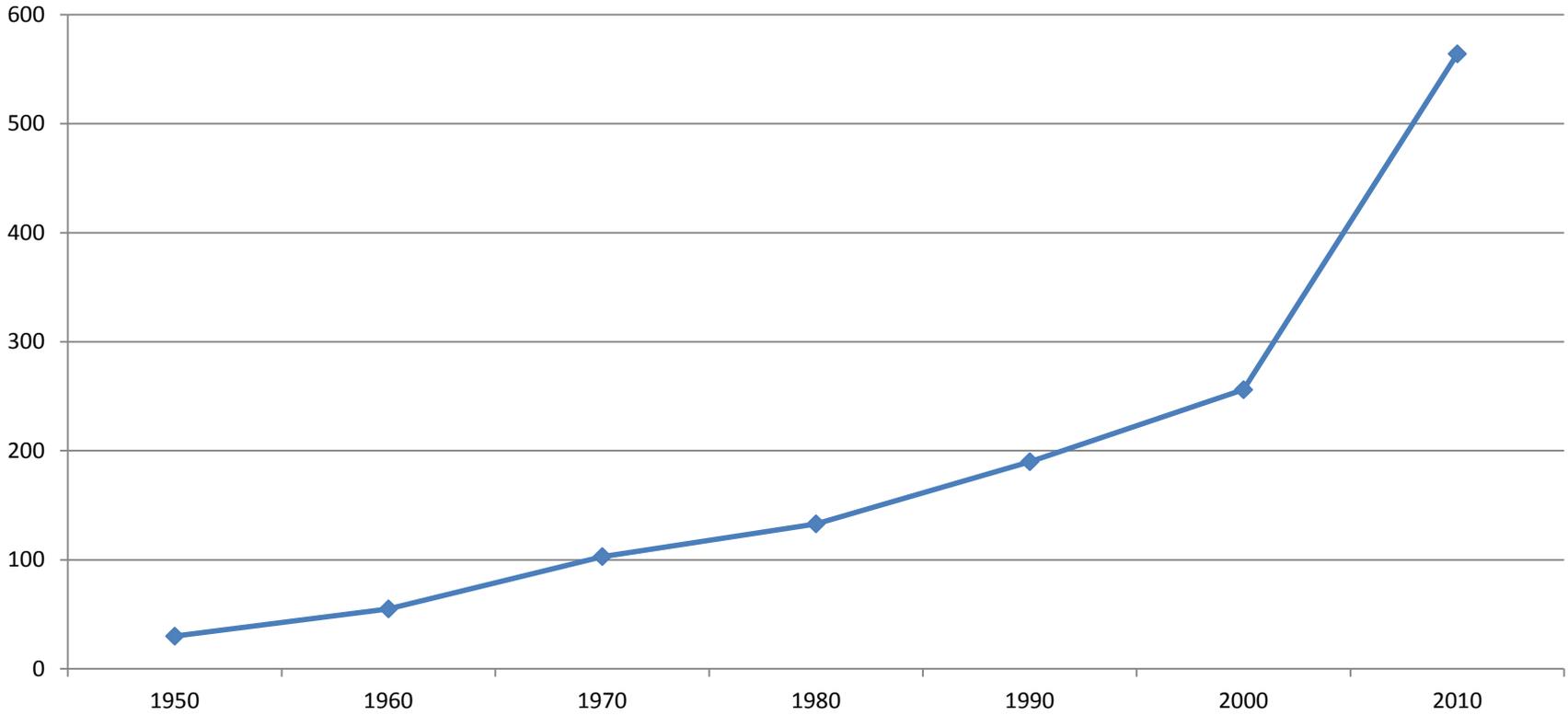
Challenges for Wikimedia growth in Indian Languages – Efforts of CIS-A2K

- Dearth of quality content available in digital format. [Digitization of Encyclopaedias and re-releasing under CC; Konkani Encyclopaedias with Goa University partnership; Kannada Encyclopedias]

- Different standards/formats/generations (Gov.in/DLI)

- Inadequate knowledge sharing culture on the internet (e.g. bit torrent)

- Insularity – Niche Communities (TA WP vs. OR WP) [Focus on diversifying users by bringing thematic experts]
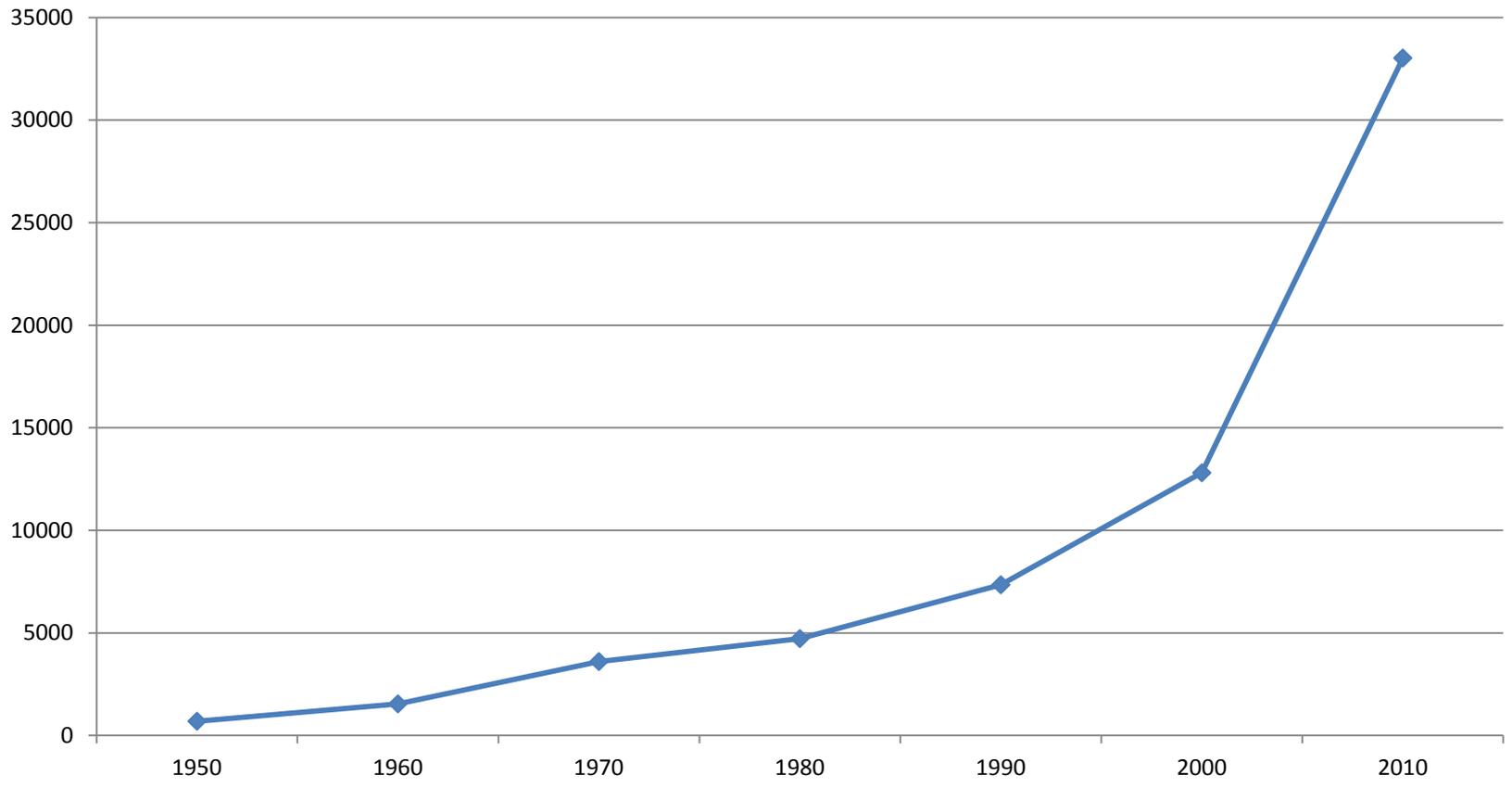
Opportunities

- Leveraging the three-language policy of Education (e.g. Kerala model) – creating future Wikipedians/knowledge producers in Indian languages

- Boom in Higher Education
  - Harnessing student time
  - Integrating Wikipedia into the pedagogy than add-on [Christ University Partnership]

✉ vishnu@cis-india.org   vishnuvardhan_t

Growth of Indian Universities

vishnu@cis-india.org   vishnuvardhan_t

Growth of Colleges in India

Opportunities

- Creating multi-lingual thematic projects [Lilavathi's Daughters]
  - Creating a pool of resources
  - Structured project
  - Cross language community efforts
  - Building a sense of healthy competition
- Sharing the periodic metrics and analyses with the community
- More on other activities of CIS-A2K www.bitly.com/cis-a2k

Some Questions?

What happens to diversity of the 700 odd Indian languages?

Do we need to put energies in bringing them onto Wikimedia projects?

Language diversity preservation needn't be tied to encyclopedic knowledge production.

Is there a way Wikimedia movement can help preserve the (Indian) Language diversity?

✉ vishnu@cis-india.org   vishnuvardhan_t

Digital technology and the possibility of having languages in their original form – the spoken word.

Post the print technology writing was important but it was only a short interlude in the evolution of humans and languages.

✉ vishnu@cis-india.org   vishnuvardhan_t

WikiSpeech

Learning from Grierson's LSI (What to collect)

Ganesh Devy's PLSI (How to collect)

CGNet Swara (Potential of Mobile)

Pad.ma (visulization and meta-data)

Media-wiki and the

Wiki model

All the sources used in this presentation are duly acknowledged at the bottom of each slide. Sources under copyright used in this presentation are used as per the "fair-use" clause of Indian Copyright Act.

This presentation is licensed under CC-BY-SA 3.0

T. Vishnu Vardhan

Prg. Director Access to Knowledge, Centre for Internet & Society

✉ vishnu@cis-india.org   🐦 vishnuvardhan_t