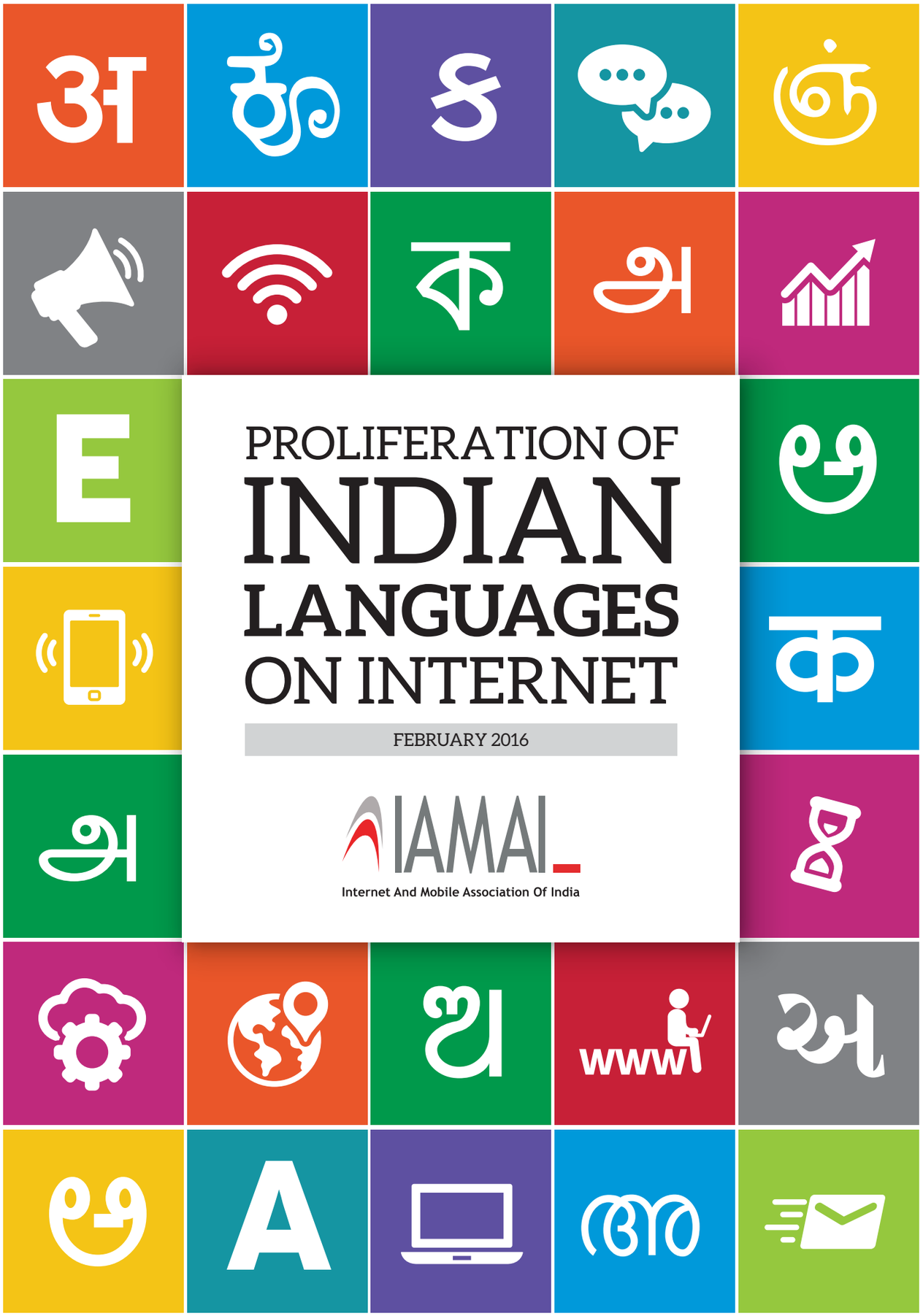


PROLIFERATION OF INDIAN LANGUAGES ON INTERNET

FEBRUARY 2016

IAMAI

Internet And Mobile Association Of India

Foreword

Digital penetration in India has been phenomenal in recent times; with India boasting the second largest internet user base in the world. However, this is just the tip of the iceberg, as a large sizeable population still lies outside the ambit of the digital connect. While growth of the digital infrastructure is one of the focus areas in this regard, there are other factors that need to be taken into consideration for achieving greater digital and mobile connectivity.

The issue of language is extremely critical, which remains largely unaddressed and therefore an impediment to the growth of Internet in India. For a culturally diverse country like India, it is not one language which can connect the length and breadth of this vast country. Digitisation will help in addressing many of the socio-economic challenges of the country by providing a fast, efficient and cost effective method of delivery of certain services. Increased digital penetration will spur the growth of numerous associated industries, leading to employment, growth and economic well-being. However, to realise this potential, the digital & mobile content industry and mobile technologies has to reach out to each and every Indian in the language they are most comfortable in.

As an industry association representing the digital and mobile content industry, it has always been our endeavour to help the industry's growth by providing studied insights on the different issues that challenge the ecosystem. This report is one such effort on our part.

The report would not have been possible without the active help offered by numerous institutions, organisations and individuals.

I would like to thank all those who have given their time and offered assistance to understand the subject matter and address the survey undertaken. Through this report, we have skimmed the surface of a subject so vast that it requires much more sustained effort to be able to understand the importance and be able to arrive at definitive recommendations which may be treated as industry standards to be followed. This volume is a fore-runner to an exhaustive study which we would endeavour to undertake in the coming year.

I would like to thank Feedback Business Consulting for the effort they put in to conduct the survey.

Every care has been taken to check the facts and arguments made out in this report and we stand by them. Despite our best efforts, it may be possible that there exist some errors in this report. On behalf of IAMAI, I take full responsibility for them.

Subho Ray
President, IAMAI

Table of Contents

Introduction 04

The Industry Ecosystem 08

Challenges In Content Generation 10

Technological Challenges & Limitations 12

Challenges In End User Searchability 17

Conclusion 19

Annexe 23

Introduction

The rapid spread of internet in India over the last few decades has been phenomenal. Today, India has the third largest internet user base in the world, out of which more than 50% are mobile internet users. The total number of internet users in India rose from 300 million in December 2014 to 402 million by December 2015. The number is expected to reach 462 million by June 2016, which would make India the second largest internet user base, overtaking USA. Mobile Internet user base in urban India has grown 65 per cent over last year to reach 306 million in December 2015, while the rural user base surged 99 per cent to 87 million by December, 2015. This is expected to grow to 262 million (urban) and 109 million (rural), respectively by June, 2016 .

Nonetheless, the internet penetration in India is still very low at around 21% approximately, as compared to 46% in China, 53.4% in Brazil or 86.8% in USA. This provides a huge potential for future expansion, and all estimates suggest that the present growth process is yet to peak. The associated business potential for such a market is huge and subsequently the Indian Internet sector is drawing global attention. It is expected that much of this connectivity will be achieved via mobile phones, as per recent trends.

Internet penetration via mobiles is expected to bring in a social revolution in rural India. Some of the important fields in which digitisation can be of immense importance are:

- Inclusive Banking: financial inclusion via the spread of formal banking is one of the prime targets of social empowerment; and the Pradhan Mantri Jan Dhan Yojana is a proactive step in that regard. Access to bank details via a simple sms is one of the tools adopted in the project. This is but a small step; and the experiences of Kenya or Uganda, where mobile money (or mobile wallet) has helped bridge the social gap serve as benchmarks for how developing countries can utilise a simple feature-phone to solve the challenge of financial inclusion.
- M-governance: is the logical extension of e-governance. Digitisation allows government services to reach out to every individual and access to services via the mobile phone makes availing the services much more convenient. This is the best possible way of empowering the marginalised sections of society.
- Weather prediction and Disaster Management: one of the prime objectives of any disaster management programme is early warning under preventive actions. Digitisation allows for fast and effective warning mechanism that can reach out to each individual via the mobile phone. Effective weather forecasts can help the farmers prepare against the vagaries of nature, thereby helping prevent crop loss or damages by considerable levels.

The socio-economic benefits of a robust digital infrastructure have not escaped the attention of the State authorities, and the present Indian Government has announced a slew of projects to bolster growth of internet penetration in India. The Union Budget of 2014-15 talked about provisions of INR 500 Crore under the National Rural Internet and Technology Mission that will focus on building the requisite infrastructure for digital reach out to the rural areas in India.

Irrespective of the Inertia in the system to accept change, the sheer size of the population being directly affected (50 Cr or more than one-third of the population) will ensure that this gap will not remain unresolved for long. However, one major impediment is the issue of language of communication.

Literacy rate in India was 74.04% as per Census 2011, with urban literacy rates of 86% and rural literacy of 71%. As a measure to address the challenge of illiteracy since attaining political independence in 1947, many of the states resorted to promoting education in their respective local languages. Even though English, as a language, has a strong aspirational value attached to it, the real spread of the language has been limited given the weak base of overall literacy levels.

One of the crucial factors that can make or break the success of deeper internet penetration in India is the issue of Local language Content. Much of present day internet content is in English. Even though English as a language is the most widely accepted language across the states, its penetration is largely restricted to the urban and semi urban regions.

Consequently, the spread of internet in India is bound to face the hurdle of the language barrier once it tries to reach out to the Indian heartland. What complicates the matter further is the fact that India has twenty two distinctly separate languages, and over one thousand six hundred dialects within each of them.

Accessing the internet in local language broadly has two challenges. The first challenge is generating content in local languages and popularising such content for broader adoption. The second challenge is on the technical front, pertaining to availability of Indian scripts for generating digital content.

Government Initiatives in Enabling Local Language

The Indian Standard Code for Information Interchange (ISCII) was developed by a standardization committee under the Department of Electronics during 1986-88, and adopted by the Bureau of Indian Standards (BIS) in 1991. The ISCII was a character code for Indian languages originating from the Brahmi script, and was one of the first attempts towards developing Indic fonts for the internet.

Presently, the Technology Development for Indian Languages (TDIL) programme under the Department of Electronics & Information Technology (DeitY), Ministry of Communication & Information Technology (MC&IT), Government of India, has been working on developing tools and resources for enabling proliferation of ICT in Indian languages. The Web Standardisation Initiative (WSI) by TDIL seeks to enable all Web related standards with 22 Indian languages so that we can achieve seamless Web for every Indian.

The Government of India is also a member of Unicode Consortium, a non-profit corporation at a global level devoted to developing, maintaining, and promoting software internationalization standards and data, particularly the Unicode Standard, which specifies the representation of text in all modern software products and standards. Indian languages are today a part of Unicode Standards, and can be used via standardised keyboards (an essential prerequisite for enabling universal accessibility for the languages).

Content in local languages are generated in India by different agencies like private online publishers and even government agencies in varying degrees across languages. Global internet majors like Facebook, Wikipedia, Google among others provide content and support for many Indian languages. However, the availability of local language content is extremely low, for e.g.: the Sanskrit Wikipedia has 11,000 articles; the German Wikipedia 1.79 million. The statistics for Hindi, one of the world's most common languages with more than 422 million speakers, are no better - there are only 22,000 Wiki pages. Even Estonia, a small European country with a population of just 1.5 million, has more Wiki entries at 55,000 pages. No Indian language, in fact, finds a place in the top 10 global languages used on the Internet, even as India is poised to become the country with maximum internet users in the world.

Modern devices like laptops and mobile handsets today offer local language supports in many Indian languages. Manufacturers are working on enabling end users to use local languages in their daily digital activities.

While the base has been laid for local language content, a lot still needs to be done to provide the level of local language interaction required for the levels of digital penetration envisaged for the near future.

Technology Development for Indian Languages (TDIL) initiatives

Some of the initiatives of TDIL for promoting local language content development are as follows:

- National Standard in script keyboard was designed in 1991 and now it’s being modified with Unicode 6.2.
- TDIL is currently developing apps like handwriting based input mechanism which will initially support 10 languages. This app will be helpful in sending and receiving SMS
- Developed text to speech tools which would be available for Android devices
- Currently developing Machine Translation systems from English to the 8 major languages

- Using Optical Character Recognition to digitize content across the various languages
- Developing speech recognition software for quick access to online data like agricultural commodity prices, weather conditions, etc.

The English language still accounts for 56 per cent of the content on the worldwide Web, while Indian languages account for less than 0.1 per cent. The anomaly here is that the Internet in India is predominantly English. However, there is high potential for regional language content. In the last year alone, Hindi content on the web has grown by about 94 per cent, whereas English content has grown only at 19 per cent.

What explains the need for web content in local languages is the growing number of Internet users in the country and the near saturation of English-speaking population already online. The number of web users will only go up significantly if Indians can access the Internet in their local languages. According to some experts, increasing the availability of local language content would also help in increasing the internet penetration by a further 24% as more consumers would be able to access the relevant content online.

There is also significant scope for the various stakeholders to benefit financially while addressing the needs of the local language content consumers. According to industry experts, if the Indian language book publishing industry moves online, it can create a digital opportunity worth nearly $7 billion for both the content providers and technology players. Further, it is also much more cost efficient to reach out to the regional language consumers in tier-2 and tier-3 cities through online sources than through traditional media- TV and print.

Private companies' efforts towards building the local language ecosystem

- MoFirst Solutions has developed a "regional" mobile phone brand, called Firstouch. Using the swipe technology, a user can write a text message, an instant message on WhatsApp or even an e-mail in Gujarati and swipe. The message gets delivered in English. Alternatively, messages received in English can be swiped to translate into Gujarati. MoFirst has also tied up with Micromax for providing the local language keyboard software in their phones
- Process Nine Technologies works with mobile phone makers such as Lava, besides e-commerce companies, to localise the device interface and translate content from English to other languages
- Pune-based LinguaNext is helping many public sector banks make the transition from English websites and apps to local language ones. The company's products translate enterprise applications into many languages
- Many other mobile handset manufacturers like Samsung, HTC, Apple & Karbonn are providing support for local languages in their phones
- Google has launched the Indian Language Internet Alliance, a group of content and technology companies, to "accelerate building of Indic language content" online
- Over 20 start-ups are working on content discovery, localising content, Indian language keyboards, fonts, user interfaces and speech recognition

Government Initiatives in expanding Social sector reach

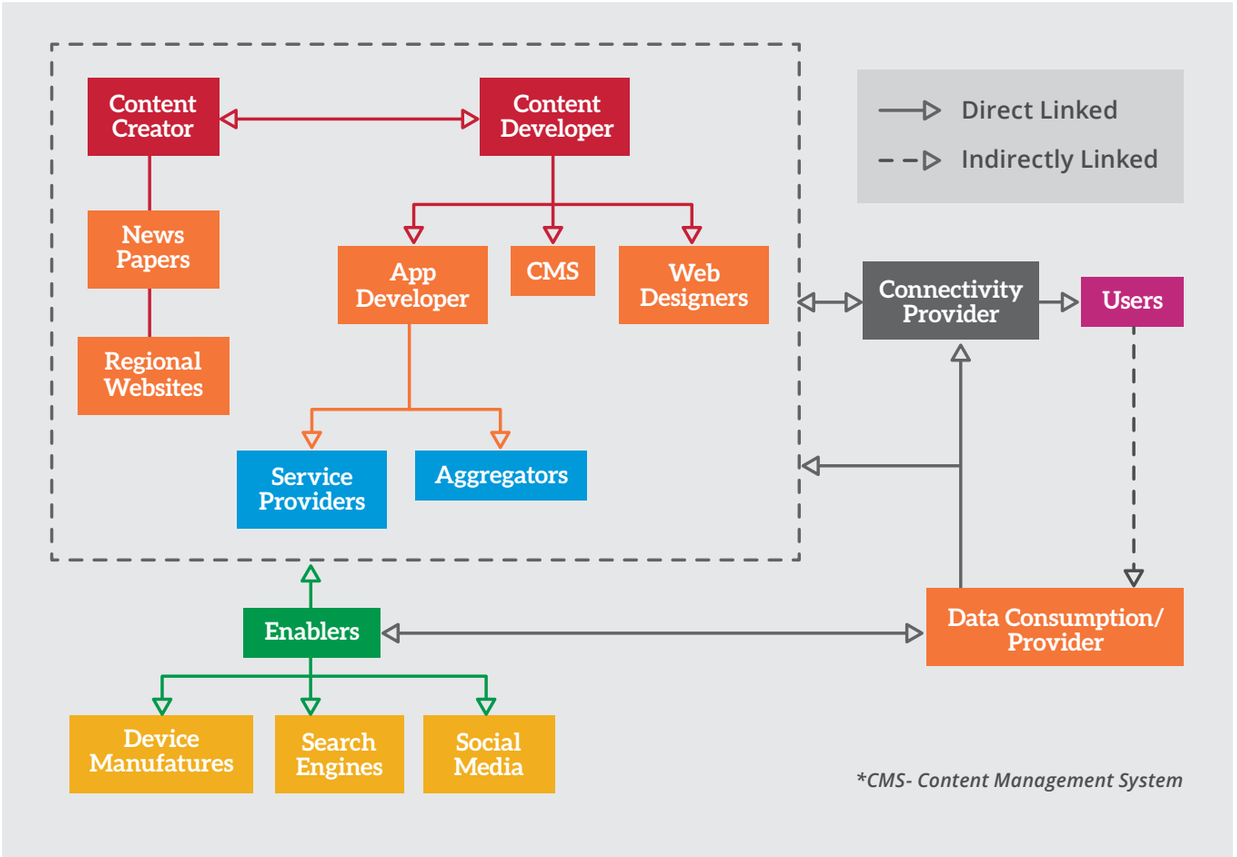
Vikaspedia is an online information guide launched by the Government of India. The website was implemented by C-DAC Hyderabad and is run by the Department of Electronics and Information Technology, Ministry of Communications and Information Technology. It is built as a portal for the social sectors, and offers information in the domains of Agriculture, Health, Education, Social Welfare, Energy and e-Governance, in 23 languages.

This report seeks to provide a comprehensive overview of the present levels of proliferation of Indian languages on the internet. This includes:

- Identifying the key players and stakeholders in the ecosystem
- Mapping the various developments that have taken place and also the initiatives being taken by the stakeholder for the future, and
- Recognizing the challenges and bottlenecks restricting further proliferation

The report starts with a brief overview of the industry ecosystem of local language content generation, followed by a brief outline of the report. The broad headings of the report pertain to the various aspects of local language content on the internet. It starts with content generation and goes on to discuss the various technical challenges involved, down to the end-user discovery of the content. The final conclusion seeks to sum up the findings and suggest a possible way forward that can help resolve the bottlenecks.

The Industry Ecosystem



A broad understanding of the ecosystem is required to better understand the challenges that are espoused in this study. The key entities in the ecosystem and the role played by them are as given below:

♦ Content Creators

Local language content for the internet starts with the **Content Creators**. These are the agencies that generate material in local Indian languages that are then published on the web. The most prolific content creators are the local language newspapers and regional websites. With the popularity of the internet, many of the major newspapers in various states catering to their own local languages have gone online in their attempt to reach out to a wider diaspora. Besides newspapers, certain local languages have other prominent Content Creators that provide different types of content. However, majority of the local language content that is currently available primarily covers News or Entertainment items. In recent times, the drive for e-governance has resulted is different State governments actively stepping in the field of generating local language contents in the digital sphere.

♦ Content Developer

While Content Creators work in providing the actual content, the digitisation of the content is handled by **Content Developers**. Content Developers provide the technological bridge for Content Creators to publish their content in the digital space and make them accessible for end users. There are different levels or layers of Content Developers.

- **App developers:** At the first level there are App developers who help the content creators in developing and testing applications across platforms to make sure that the content is rendered properly, without glitches. Due to the overall lack of local language content online, there is a general dearth of app developers for local languages. Then there are service providers who closely work with the content creators and design the

customized application according to their needs. Finally there are the aggregators who source content from different sources and develop their own applications.

- **Web Designer:** At the second level, we have web designers who help the Content Creators in designing the web pages in order to make sure that the content is displayed appropriately across different types of browsers, operating systems, etc.
- **Content Management System (CMS):** CMS Providers help Content Creators to type, publish and store their content on the web in a systematic and efficient manner.

♦ Enablers

The Enablers ensure that the content thus published on the virtual space reaches the end users. The category encompasses agents engaged in a wide range of activity.

- **Device manufacturer:** The manufacturers of computers, mobile phones, keyboards, chips, and related hardware whose role is to make sure that their respective devices provide seamless support to all the major languages by default. Mobile handset manufacturers play a key role in promoting the use of local languages since most new internet users are coming up on mobiles and not on laptops/ desktops. The Indian mobile handset market is the third largest in the world, after China and USA. Samsung and Micromax are the key players with about 60% share of the market.
- **Search Engines:** Used by most to surf the web, search engines are often the first port of call for most users in the virtual space. Hence, they have a big role in ensuring that the end users are able to search for content in any language of their preference in a way that is user friendly; even if the user is not well versed with English and maybe new to the internet.
- **Social Media:** The advent of social networking sites has today provided the space for individual users to create and share local language content online. Social Media thus has a big role in facilitating local language content usage in the near future. Facebook and Twitter are the key social media websites in the Indian market for creating and sharing content. Facebook has over 100 million active users in India at present.

♦ Users

At the end of the spectrum are the users who consume the local content. Currently there are approximately 300+ million users who are fairly comfortable with English and know how to navigate the internet. There is also a large pool of potential Users who presently face challenges in using internet and are expected to connect to the digital world in the near future, given the rapid expansion of mobile usage and the Government of India's stated vision of connecting the unconnected.

♦ Consumption data/ analytics provider

An important agent in the ecosystem is the data analytics provider. They offer analysis on the online content consumption for various websites, disaggregated at various levels and parameters, which is an essential tool for measuring the extent of content consumption.

♦ Connectivity provider

They provide connectivity (Broadband, Telecom, etc.) to the various stakeholders in order to enable them to interact with each other seamlessly

The process of content creation and content development can be clubbed together to define the category of **Online Publishers**. Publishers are agencies that 'publish' contents on a website, generated by **Content Creators** and **Content Developers**. In many cases, publishers are original **Content Creators** who resort to third-party **Content Developers** for help. In many cases, certain aspects of Content Development are done in-house by the **Publishers** (for example, most major regional newspapers have their in-house IT department who look after the daily affairs of their website CMS, even if the initial designing of the site may have been outsourced).

Scope of This Study

For simplicity, Publishers are taken as the first port of call for studying the problems of Content Creators and associated content development issues. The issues raised by Publishers are mitigated by the Technology providers like Device Manufacturers, Content and App developers, etc. Feedbacks on technical issues are taken from Content Developers and Enablers, depending on their fields of expertise.

The study is based on an extensive survey of the key players of the local language ecosystem, undertaken by Feedback Business Consulting, at the behest of IAMAI. The detailed list of those interviewed is provided in the Annexure.

Challenges In Content Generation

♦ Present Scenario

- Most of the media companies have in-house team for content creation i.e. news collection, story making & translation, and uploading on websites. They also invest in training manpower for overcoming the other challenges pertaining to content availability.
- Currently, local content availability is restricted largely to news agencies as other forms/sources of content generation are extremely limited. User generated local content is still very low, as adoption of local language usage in social media and other such fields are still restricted.
- Central Government department websites are predominantly in English. Limited amount of these contents are available in Hindi as digitisation of Hindi content is still limited. Many State Governments have started offering their websites in the respective local languages. Efforts in promoting e-governance hinge strongly on the availability of such services online in the local language for greater acceptance and adoption. However, digitising content in local language is still restricted given the technological challenges that still exist.

♦ Challenges faced

- Barring Hindi, no other local language covered under the survey has a reliable agency providing news in their local language. Thus, invariably, the publishers need to invest considerable manpower that adds to their expenses. This also generates challenges in terms of training employees to collect, translate and upload data, availability of audio video content, breaking live news as fast as English, etc.
- Most internet sites provide free access and companies earn revenues out of advertisements, which in turn is dependent on the number of users accessing the site. This in turn gives rise to different set of challenges for some local language newspapers. Newspapers like Bengali and Tamil are facing serious copyright issues from neighboring countries Bangladesh and Sri Lanka who copy and publish their news, thus reducing the viewers on their website.
- Much of the internet activities, like e-commerce, entertainment etc. are still largely English centric in India. As a result, much of the user interactions in these fields are restricted by language.

  For example, Facebook, one of the most popular social networking sites in the world, does not yet have easy input tools for Indian languages, which makes engaging in local languages a major challenge. Even though there are many users who log into Facebook via the Hindi website, their activities are severely restricted as most activities on the site are English-centric.

  Till now Facebook supports 8 major Indian languages, while Twitter presently only supports Hindi. Google is one major internet agency that is working most proactively in supporting local languages in their searches and other activities.

  Unless these websites provide greater support for local languages, it will be difficult for users to use local languages in their online activities.
- A major restrictive factor in availability of local language content is the low levels of public awareness about the fact that local languages are supported in the Internet. For example, Facebook initially did not have the option to click on “Hindi” as a language option upfront on its landing page until recently, despite having the facility. The key reason for this was that most of the end users are not aware of how to navigate websites; they do not understand drop down lists, etc. Hence, only after the Hindi language option was highlighted on the landing page did Hindi registrations really take off on Facebook. Wikimedia developed a JavaScript system for typing in Indian languages, but it did not receive the kind of response they expected. Thus, availability and awareness are both crucial factors for promoting local language content availability in the coming days.
- Even the E-commerce websites, for whom a major chunk of the sales are coming from the Tier-2 and Tier-3 cities, are primarily in English as of now. Only Snapdeal has launched the website in 12 major Indian languages, while the other companies are still working on it.

♦ Action Areas / Recommendations

Some of the following initiatives can help increase public awareness and engagement in local language content.

- Development of websites in local languages and providing a user friendly interface for local language user can be an integral part of e-governance initiatives by the State Governments. This will not only help spread awareness, but also help a larger section of the population avail the services.
- Media companies should adopt digital media as a separate entity from print media and actively expand their e-presence. As of now, the former is seen as merely an extension of the latter.
- Increased spending on online advertisements in local languages would help the companies generate higher revenues through online content in local language and provide incentives for more publishers to venture in local language content.
- Digitalization and online release of the various local language archives and books in various libraries and universities across the country can be undertaken to increase local content availability. The National Digital Library initiative by the DeitY in association with the Indian Institute of Science (IISc) and other partner organizations was a good step in this regard. Unfortunately the project failed to gain the kind of traction it needed to.

  The State governments can join the National Digital Library project or initiate such content pools which can then be shared for free consumption. Certain agencies like Apex CoVantage provide technological solutions for easy and efficient translations of documents in English into different local languages.

  CBSE board has digitalized all their study material and made available to the students at minimal cost. State Education Boards can undertake similar activities to ensure ease of access of educational tools in their respective States. This in turn can encourage e-learning initiatives that can have a wider reach in remote areas.

  Such initiatives need to be advertised heavily to the target audience to ensure greater adoption of these services.
- Social media companies need to support local languages and promote their availability. The interface has to be user friendly so as to provide ease of interaction for the users.

These initiatives would help in creating a buzz around the availability of regional language content online and encourage more people to access the internet in local Indian languages. State Government can take such initiatives or incentivize them for private agencies to engage in such projects. Creating a pool of local language content can encourage individuals to generate their own contents in their own languages, thereby helping promote adoption of those languages in the digital sphere.

Technological Challenges & Limitations

Unfortunately, a major restrictive factor in generating local language content is the various technological barriers that exist currently. The various aspects of it are covered in the following sections.

Problems Related to Local Language Fonts and Keyboards

Proportion of respondents facing challenges with respect to rendering of fonts across different types of devices, like mobiles, tablets, computers, etc. (Statistics are based on the sample covered)

Language	Newspaper	App Developer
Hindi	25%	17%
Marathi	33%	17%
Malayalam	50%	50%
Gujarati	50%	50%
Kannada	50%	50%
Bengali	50%	50%

Often, the first stumbling block for publishers venturing in local language content generation is the problem of rendering the contents in the script of the language. The primary building block is the font of the script.

The challenge is not only in having digital fonts available, but also in how to display it to the end user in the digital space. Rendition of fonts on the screen of a user's device depends on

(i) The type of device (laptop, tablet, mobile phone, etc.);

(ii) The subsequent software platform it operates on.

This is often a separate challenge for developers.

♦ Present Scenario

- Several publishers such as Ananda Bazar Patrika (ABP), Sakaal Times, Manorama Malayalam and Deepika have developed proprietary fonts and continue to use them because Unicode was not supported well at that time. While most of the present age publishers are transiting to 'Unicode' in order to follow standardization, ABP refuses to shift because of certain specific issues. Also, many Kannada websites cannot be searched right now since they are not Unicode compliant.
- Currently, most of the regional language news sites are using 'In Script' based keyboard where several different font layouts (i.e.: mapping of English keys to local language keys) are available.
- Unicode is compatible across platforms and hence publishers are encountering no major issues in rendering the data on websites accessed by computers. However, there are few issues encountered while rendering the data on the mobile devices.
- Inclusion of Indian Language fonts and keyboards in mobile handsets depends on the discretion of the hand-set manufacturer. The problem is more acute for feature phones (which still forms the bulk in the non-urban economically weaker areas) as none has an easy-to-use input method for any other language other than English.

Indian manufacturers providing all language support need government support in forms of subsidy on component imports to encourage proliferation: *Vineet Taneja, Micromax*

- Facebook and Google are both developing "Voice to Text" software for different languages and the industry feels this could well turn out to be the next killer application. Once in place, it has the potential to drastically increase user generated content which would supplement content provided by the publishers. Wikimedia had earlier developed web-fonts for Indian languages in their websites. Unfortunately, it created performance issues for the site because of which it was turned off as default.

♦ Challenges faced

Percentage of respondents reporting challenges in working with complicated input mechanisms and related problems (Statistics are based on the sample covered)

Language	Newspaper	App Developer	Device Manufacturer
Hindi	25%	50%	0%
Marathi	100%	100%	100%
Malayalam	100%	100%	100%
Gujarati	100%	100%	100%
Kannada	100%	100%	100%
Bengali	100%	100%	100%

- Though Unicode, on the whole, is compatible across platforms and helpful to the publishers in increasing their viewership, it has its limitations that create certain impediments. Presently there are issues with visual appeal of Unicode. Moreover, having more than one standard and different font styles would also enable the users to choose the one that suits their preferences.
- Space limitations, missing characters, complicated typing mechanisms, etc are some of the challenges of generating content in local language fonts. The complicated typing mechanism makes it difficult for amateur content developers to use local language keyboards.
- With complex characters such as compound letters and zero width, some challenges might crop up especially on mobile devices.
- Screen sizes changes from one phone to another and from one tablet to another with different resolutions. Fonts often do not render well across devices as the font developed for a bigger size device is not suitable for lower size device. For e.g. Google has developed proprietary font 'NOTO FONT' whose compatibility is only checked on high end smart phones. However, it should be more focussed on low end mobile devices which are used by masses.

♦ Action Areas / Recommendations

Some initiatives that can help in developing user friendly local language scripts are

- While the Government agencies have already developed a standard font layout (i.e.: mapping of the English keys to the local language keys) for each language, it is yet to be adopted as an industry benchmark. As the market is gravitating towards 4", 4.5" and 5" for phones and 8" and 10" for tablets, there has to be some standardisation followed for development of fonts. This needs to be done based on the In Script and other popular font layouts that are easy enough for the end users to learn quickly and type faster. The fonts should be made open source so that any bugs in the system can be worked on by the other players in the language infrastructure development community.
- Government has to mandate Indian languages over mobile devices. Technology provider/device manufacturer have to manufacture localized devices and apps thus making it user friendly for both content developer and users to post in regional languages. Subsequently, all the local language fonts should be made available across devices - computers, tablets, smartphones, feature phones and the browsers straight out of the box, so that the end users do not have to look around to download the same.

  This is crucial, as much of the spread of internet connectivity in recent times is via mobile phones, and this trend is expected to continue. If local language usage is to be popularised, one has to ensure that the devices carry provisions of such languages. This will encourage greater adoption of such devices, as well as pique interest amongst users to generate local language contents.

Other Input and Output Challenges: Tools, website/app Designing and related issues

While fonts are just a specific technical aspect, local language content development requires a set of tools and software for generating digital content. Tools like digital dictionary, predictive text, etc are more readily available for English but are hard to come by for any of the Indian languages.

Designing a website is another technical aspect. In today's world, many entities have to maintain separate pages for mobile and computers, depending on the type of technology or software used by each in accessing the net. Designs of apps differ from websites depending on certain technical parameters.

♦ Present Scenario

Table 1: Details of in-built local language support currently provided by the major device manufacturers

Device Manufacturer	No. of Indian languages covered	Device type	Type of keyboard	Handset Price range (INR)	Software development
Micromax	10	Mobile phone	Touch pad, Keypad	3,000 to 30,000	Partnered with MoFirst
Samsung	9	Mobile phone	Touch pad	5,000 to 50,000	In-house
HTC	10	Mobile phone	Touch pad	5,000 to 50,000	In-house
Apple	8	Mobile phone	Touch pad	20,000 to 80,000	In-house
Lava	21	Mobile phone	Touch pad	Less than INR 5,000	In-house
Karbonn	21	Mobile phone	Touch pad	Less than INR 5,000	In-house

Table 2: Details of tools used (dictionary, spell check & translation) by the Newspaper companies

Language of the Newspaper	Types of tools required	Key supplier/ resource used (Based on the sample covered)
Hindi	Dictionary	Various websites like Shabdkosh, Hindkhoj, etc.
	Spell check	Manual
	Translation	• Google translate • Manual
Marathi	Dictionary	• Some websites like Shabdkosh • Manual
	Spell check	Manual
	Translation	• Google translate • Manual
Malayalam	Dictionary	Manual
	Spell check	Manual
	Translation	• Google translate • Manual
Gujarati	Dictionary	Manual
	Spell check	Manual
	Translation	• Google translate • Manual
Kannada	Dictionary	Manual
	Spell check	Manual
	Translation	• Google translate • Manual
Bengali	Dictionary	Manual
	Spell check	Manual
	Translation	• Google translate • Manual

- There are quite a few publishers engaged in converting content from English websites and apps into regional languages, and have developed their own proprietary tools for dictionary, spell & grammar check, translation/ transliteration, etc. Others are using tools from external agencies to check for the basic spelling and grammatical errors.
- For individual users, online availability of such tools or services is yet to take shape, thereby restricting individual user activities. Agencies like Apex CoVantage have certain software that helps create content in different languages. However, such services are yet to gain traction given low user awareness and content developer base in the different Indian languages. Other agencies too have developed similar tools.

- Most content creators do not feel the need for any external support or the need for any industry wide standards for site/ app layout as they would like to display content in their own unique way. They prefer to work on it themselves or outsource it to a technology company.
- In this regard, many of them have developed in-house "Responsive Design" units, which take care of the differences in the way the content is displayed across different screen sizes. Here again, they face the problem of building interactivity into the content.
- Reverie Technologies (Bangalore) has created local caches at their customers' servers hence reducing the number of layers that the content has to go through to load on the end users devices.

♦ Challenges faced

- For a self-generated content, the content creators have to manually type the content and check for errors. The same holds good for external content as well, especially when sourced from a regional content provider. Here, they have to modify the content according to their requirement.
- The tools (dictionary, spell check, grammar check, translation & transliteration) developed in-house by publishers are usually very basic and do not serve the purpose satisfactorily. Thus, these tools demand a lot of manual intervention. On the other hand, the tools available in the market, like Google Translate, online dictionaries, etc. are helpful but only to some extent. For e.g.: One cannot simply put an English sentence in Google Translate and expect it to provide output which can directly be used by the publishers, since many of the words are not translated properly (eg: "13 MP camera" translates into " 13 saansad camera " by Google Translate). As a result, the publishers have to use such tools one word at a time and use their judgement and skill to make the tools work for them. This process requires a lot of time and effort, and is expensive.
- The major issue when it comes to developing a web page/ app in regional language is its size. Since regional language fonts are complex in comparison to English, such a webpage are often 'heavier'. This limits page size and it takes more time for a regional language web page to load.
- Limitations mentioned above often restrict site versatility that ultimately affects the end-user's experience.

♦ Action Areas / Recommendations

- The Government agencies like TDIL, who are active members of the UNICODE Consortium, are currently working in the development of basic tools such as dictionary, spell check and grammar check to make it easier for content creators (both companies and end users) to create and edit content. The process has to be expedited with a definite timeline which must be followed rigorously.
- Such tools should be programmed into the chip and provided as in-built offering along with fonts. It will make content creators job easier in creating and editing the content and enable individual users to start generating their own content.
- The industry should support and encourage the development of fonts and layouts that make the local language content lighter and hence faster to load than it is at present. This would help in improving the end users experience and subsequently increase its usage.

Challenges In End User Searchability

While publishing online content is just one aspect of local language propagation, discovery of the content by the end user is the other half of the picture. A big challenge in this aspect is the fact that much of internet searching is English based and regional languages are yet to attain a basic level of ease and popularity till date.

♦ Present Scenario and Challenges Faced

- Out of the current 400 million+ internet user in India, only a very small percentage of users are aware of the existence of regional language content. Most are accustomed to English content by habit and are used to accessing English based websites for their regular activities; even for sites that do support local languages.
- Currently, a user has to perforce use English to access sites with regional content. Once the user reaches the site, he/she would aspire for high level of optimization and accuracy in terms of searchability, which in most cases may not be available in the local language. This means that users are still predominantly depended on English for navigating the web, which defeats the whole purpose of generating local language contents.

Why English is the Predominant language of Usage

Social pressures
"...Other communicate in English on internet ..how can we use vernacular.."
"...Primary use of internet is for emails and emails don't support vernacular..."

Lack of supporting apps and websites
"...Most apps don't support Kannada so we have to use them in English..."
"...Government websites are in Kannada but typing in their forms in Kannada is so difficult..."

Scientific information is mostly in English
"...All scientific words cannot be translated to vernacular...so such information has to be read in English.."

Lack of awareness regarding vernacular content
"...Due to unavailability of Odiya we use English as language for accessing internet..."

Inability to understand the vernacular language used on the net.
"...In Telugu we cannot understand the options so we go to English..."

Incorrect translation
"... There was some very different word used for the ingredient Rava which was very difficult to understand..."

Phone is used to access internet and phone does not support vernacular font
"...I Did not try Tamil font in mail because I don't know whether my mobile will support it or not..."

Unable to understand vernacular used on internet
"... When it comes with pure Tamil literal word, we cannot understand it. So if we find out English meaning, we can understand what is it ..."

Tools like vernacular dictionary and keyboard are not helpful
"...if we go there, and ask, 'porul' (meaning), it will come with their meaning....I used dictionary when I wanted to know the meaning of, 'eetheni', I wanted to know what will be the content in it, but even then I could not open it..."

(The quotes are illustrative examples from field survey of local language users)

♦ Action Areas / Recommendations

- The tools currently available for searching the local language content on the www or mobile web (for e.g.: how to change the search language) need to be popularised by public campaigning. Notwithstanding the people in metros who are aware of the existing tools, there is a need to create awareness and socialise these tools in tier 2 & 3 cities and rural areas to popularise the usage of local language content.
- Further, the Government must encourage the development of software for handwriting recognition, and voice to speech among others. This will make the end users' task easier in creating and searching the content in their preferred languages.

  For e.g. Google has launched a website www.hindiweb.com to help Hindi-speaking Internet users discover Hindi content across websites, apps, videos and blogs. Google is planning to launch similar websites for other major Indian languages such as Bengali, Tamil and Marathi, over the next 15 months.
- As a part of its "Sarva Shiksha Abhiyan", the government can educate the end users about availability and usage of the various existing tools/apps to create and search local language content online. The government must also create a separate ecosystem for local language apps. This would make it easier for the end users to discover such apps resulting in more awareness and usage.
- Furthermore, companies that have websites in several languages could be provided with the relevant location data of the end user who are accessing the site through their IP address/ mobile number, etc. This would help the companies to offer their site/ app directly in the relevant local language of the end user, instead of the end user having to go to settings and select it. Moreover, this initiative is bound to ensure a smoother experience for the end users.
- Indexing and storing data is another challenge where the industry has to come together and create self regulatory norms for the indexing of the content and the way content is stored. Voluntary adoption of 'Best Practises' can help the industry as a whole give out a positive signal to the larger mass about its seriousness to promote local language content. This can not only be a positive step for bolstering the industry's image, but can also act as a sales pitch to encourage more and more users to join the digital world.

  Systematic indexing and content storage will boost the rendering of content across different platforms and improve the searchability of the content, thereby making them more usable.

Conclusion

The local language content penetration on the Internet is increasing over the time but at a sluggish pace. The Urban penetration reached up to 43% whereas the rural penetration stood at 57% showing the higher acceptance of local language content in the Rural Internet users. It is estimated that enabling local language content on the Internet will lead to a growth of 39% in the current Internet user base. Out of this, 16% growth will come from urban usage and 75% growth from the rural users . However, the growth rate is impeded because the industry has failed to act together in a concerted manner.

Lack of Cohesive action

During the process of this study, it was increasingly realised that the biggest challenge in the field of local language proliferation is the fact that most of the sectors/agencies are working in silos. This leads to two broad set of problems:

a) Most agencies are reluctant to adopt developments made by their peers and insist on working on their own; as often the peers are market competitors. This often leads to problems of different agencies working on the same issue simultaneously. For Example, on the issue of fonts, publishing houses, technology platforms and other agencies have developed parallel set of fonts. The same applies for input tools, dictionaries, spell checks, etc, where almost every agency have developed in-house capacities, which is a highly resource intensive process.

b) Autonomous developments lead to the problem of synchronisation. Without standardisation, problems of rendering, webpage/app designing etc are bound to crop up, leading to further complications.

Revenue Generation

While a broad consensus on potential market does exist, no definite estimations are available for language specific market potentials. Developing in-house capacity for generating local language is resource intensive and hence an expensive proposition. Unless a definite revenue potential can be ascertained, the field will remain restricted to big players with deep pockets and captive markets. Market surveys are crucial to enable potential entrepreneurs have an idea of ROI so as to develop their business models.

Since most web contents are on open platforms with free access, monetisation is a challenge. Given that mobile internet is the future in communications, much of the developments going forward will in the app space, where monetisation is a major concern. Monetisation in the app sector still banks heavily on digital advertisements. However, digital advertisement in local languages is yet to pick up. On Television, irrespective of the channel's broadcasting language, on an average 4 out of 5 Ads are in local languages. Even the online channels like YouTube and online broadcasting sites are producing and disseminating the Ads in local languages. The overall digital advertising spends in India is estimated to be about INR 3,575 Crores by the end of December 2015. The proportion of Digital ads spends in the local language is approximately 5% of the entire market i.e. INR 179 Crores. With the increasing availability of digital content in the local language, this share is expected to reach close to 30% of overall digital advertising spends by the year 2020 .

Presently, the Directorate of Advertising and Visual Publicity (DAVP) rates empanelled websites in 3 Categories; Group A- having over 5 million unique visitors ((from India) per year, Group B – having 2 -5 million unique visitors (from India) per year and Group C - having less than 2 million users (from India) subject to minimum requirement. The DAVP sets advertisement rates for each category, as shown in the table below.

DAVP RATE FOR EMPANELLED WEBSITES*

"X" indicates there are no rates finalized for certain banner sizes.

Group A			
Sr. No.	Banner Size		Rate(in Rs) on CPTI Basics
1	728 x 90	Top	90
		Bottom	X
		Side	X
2	300 x 250	Top	X
		Bottom	X
		Side	120

Group B			
Sr. No.	Banner Size		Rate(in Rs) on CPTI Basics
1	728 x 90	Top	110
		Bottom	30
		Side	X
2	468 x 60	Top	35
		Bottom	X
		Side	X
3	300 x 250	Top	35
		Bottom	X
		Side	44

Group C			
Sr. No.	Banner Size		Rate(in Rs) on CPTI Basics
1	728 x 90	Top	23
		Bottom	15
		Side	23
2	468 x 60	Top	11.5
		Bottom	9.2
		Side	11.5
3	300 x 250	Top	23
		Bottom	23
		Side	23
4	234 x 60	Top	9.2
		Bottom	X
		Side	X

Since the categorisation is done as per user statistics, it is no surprise that local language websites more often than not qualify under Category C which has the lowest rates. This in turn means that such sites earn the lowest rate of revenues from adverts. The DAVP can offer special rates to local language content providers as an effort to promote development of such sites.

Differential rate for advertisement is a dampener for local language publishers and needs to be addressed immediately: *Gyan Gupta, COO, Dainik Bhaskar*

The publishers too need to understand and develop the ideal site/ app layout that are user friendly for even the most layman of a user. Once the website becomes user friendly, its usage is bound to increase. This can help them rise in category, earning higher rates per adverts.

Statistics on number of site hits and other such details help understand the type of users presently visiting certain sites. Analysis of such data can help identify the lacunae, highlight best practises and help develop an overall understanding about the popularity and shortcoming of present endeavours in promoting local language content in Indian languages.

Currently, the publishers are totally relying on Google analytics to measure the data consumption by users on their website. There is a need to standardize the source of data, which gives more details on the type of end users accessing their websites. The Audit Bureau of Circulation is said to be working on creating an industry wide data standard. This will help create a centralised source of data to analyse the consumption of all types of local language content online in India. This will also help in generating interest among the advertisers.

Content Discovery: long term solutions

The challenge of content discovery by end users is dependent upon challenges of searchability which in turn depends on indexation. While the popular search engines are developing local language searches, lack of uniformity in indexation can limit their scope. Best practises and industry benchmarks are yet to evolve which can streamline indexation. Indigenous efforts are required to resolve this problem.

Way to Go Forward

The industry needs to undertake coordinated efforts to resolve the above mentioned challenges. The Government can act as facilitator up to a certain extent. The Government can at best set some broad contours within which market dynamics will help the ecosystems to evolve over time.

The Government should take initiatives to digitalise contents in the state run libraries to encourage more user access local literature online: *Sebabrata Bannerjee, AMAR UJALA*

Immediate action items

- Increase the awareness of digitization by
  - Developing website in local languages and also incentivize popular websites to adopt various Indian languages. e-Governance websites in local languages can be a strong initiative in this regard. Currently, the landing pages in e-governance sites are in local languages but the interactive pages are in English.
  - Online spending to advertise tenders, job postings, etc. in local language can help raise digital advertisement in local languages. This would help in getting more people to access online websites and help the companies in increasing their revenues through online content. The issue of low Ad rates needs to be addressed so as to provide parity to local language publishers.
  - Government can accelerate the National Digital Library programme by devising incentives for private agencies to digitize various local languages archives and books that are currently lying in various libraries and universities across the country in user searchable format.

- Developments in UNICODE, fonts and layouts that make the local language content lighter and hence faster to load needs to be pursued.
- There exists a market for basic tools like dictionary, spell check and grammar check in order to make it easier for content creators, both companies and end users, to create and edit content. The industry needs to capitalize on this need and help set standards and benchmarks.
- The Government should help in bringing all the various organizations working in silos together to provide a single end- to end solution to the content creators. At present, all the important local language support providers work independently where some organisations provide translation tools, some provide keyboards, some provide data on online consumption, etc. This makes it very difficult for the content creators as they have to work with a number of stakeholders to get the work done. Instead, their job would be made much simpler if they have to deal with a single entity that can provide all the solutions as a package deal.
- The industry should collaborate to evolve guidelines when it comes to indexing of the content and content storage. The Government agencies can anchor such projects by bringing all stakeholders on the table and drafting a common statute.
- According to industry experts, around 60% of the news consumed in US is through various social media websites like Facebook, Twitter, etc. Government needs to incentivize the social media companies to adopt the major Indian languages in order to further improve the discoverability of local language content.
- Device manufactures must adopt the standardized font and other technological standards and also offer tools for local content development in the form of apps. This will enable greater user generated content that will enrich local language adaption.
- It is critical to have good broadband connectivity even in the remotest areas. Presently, connectivity continues to be a challenge even in the metros. Better speed of internet connectivity can help promote video content traffic in local languages. Audio-visual media is the best tool to reach the sizeable section of society who are either not literate or at best semi literate. The latter too are more comfortable in communication via audio-visual medium than through text. There is a very urgent need for the Government to intervene to make Digital India a reality.

Annexe

Stakeholders	Language	Media companies Company covered
News & Regional sites	Hindi	• India Today • Jagran • Raftaar • Navbharat Times
	Marathi	• Lokmat • Sakaal • Marathi world
	Malayalam	• Malayala Manorama • Deepika
	Gujarati	• Gujarat Samachar • Anko Dekhi
	Kannada	• Vishwa Kannada • Kannada Prabha
	Bengali	• Anand Bazar Patrika • E-Samay
	Telugu	• Eenadu • Prajasakti
	Tamil	• Dinamani
	Urdu	• Aaz Ka inqalab • Imroz-e-Hind
	Multiple languages	• OneIndia
Mobile device manufacturer	• Karbonn • Micromax • Keetronics • Apple	
Mobile Content & App developers	• Reverie Tech (Font Developers) • Wikimedia • Linguanext • Swatanthra Malayala Computing • Indicus • DailyHunt – (formerly NewsHunt)	
Others	• TDIL • Google • Facebook	

About IAMAI

The Internet and Mobile Association of India [IAMAI] is a young and vibrant association with ambitions of representing the entire gamut of digital businesses in India. It was established in 2004 by the leading online publishers, and in the last 11 years has come to effectively address the challenges facing the digital and online industry including mobile content and services, online publishing, mobile advertising, online advertising, ecommerce and mobile & digital payments among others.

Eleven years after its establishment, the association is still the only professional industry body representing the online and mobile VAS industry in India. The association is registered under the Societies Act and is a recognized charity in Maharashtra. With a membership of 190 plus Indian and MNC companies, and with offices in Delhi, Mumbai and Bengaluru, the association is well placed to work towards charting a growth path for the digital industry in India.

www.iamai.in

DESIGNED BY OSPIXELS