

# Invited Talks



# W3C India Office, its objectives and Roadmap

## Presented by :

Swaran Lata

Country Manager, W3C India Office  
6, CGO complex, Electronics Niketan  
New Delhi

E-mail : slata@mit.gov.in

1

## About W3C(World Wide Web Consortium)

- W3C is an international Standards Body which develops Standards / Best Practices / recommendations to ensure seamless web access to all
- The Vision of W3C is to achieve  
 “Web for Everyone”  
 “Web on Everything”
- W3C has so far published about 183 standards for web technology
- It works with others standards making bodies such as UNICODE, IETF, ICANN and ISO at the international level



2





## Mission of W3C India Office



**Web for All and Web on Everything..**  
 Leading the Web to its full potential..  
 वेब को भारतीय भाषाओं में सक्षम करना..

3



## About W3C India Office



- W3C India Office has been set-up at DIT under the aegis of Technology Development for Indian languages (TDIL) programme

### Objectives of TDIL programme

- Research and Development of Technology, Software Tools and Applications for Indian Languages
- Proliferation of Language Technology products and solutions
- Development of Standards for linguistic resources, tools and applications for interoperability
- Its goal is to enable all W3C Recommendations with 22 Indian languages

4





## About W3C India Office



- W3C India Office will run W3C India Portal
- Bilingual Translation and Circulation of W3C New-letters
- Establishment of National Level Special Interest Groups (SIGs) in Technology Intensive areas
- Generate national recommendations for specific Standards through Stake holder consultation

5



## Indian members of W3C



- Department of Information Technology, Government of India, New Delhi
- Comviva , Bangalore
- Indus Net Technologies , Kolkata
- Dot Com Infoway , Tamil Nadu
- Page Traffic Web-Tech (P) Ltd. , New Delhi
- Royal Website Design , Ahmedabad
- Bhrigus Software (India) Private Limited , Hyderabad
- Atal Behari Vajpayee - Indian Institute of Information Technology and Management, Gwalior
- Data Recovery Software , Ghaziabad
- Setu Software Systems, Hyderabad
- Yantra Software , Hyderabad

6






## Objectives of W3C India Office




**Objectives of W3C India Office**

- Bi-directional communication between Stake holders and W3C Consortium.
- Education and Outreach to all stake holders
- Promotion and proliferation of W3C Standards
- feedback to W3C for the existing and future standards of W3C



## Objectives of Conference



- The formal launch of W3C India Office to catalyze the promotion of W3C web standards in India
- During this conference various web technologies and aspects and challenges of promotion of W3C standards in Indian Languages would be deliberated
- Gathering feedback from participants including Industries, Government, Academia to evolve roadmap for W3C India Office





## National Advisory Committee

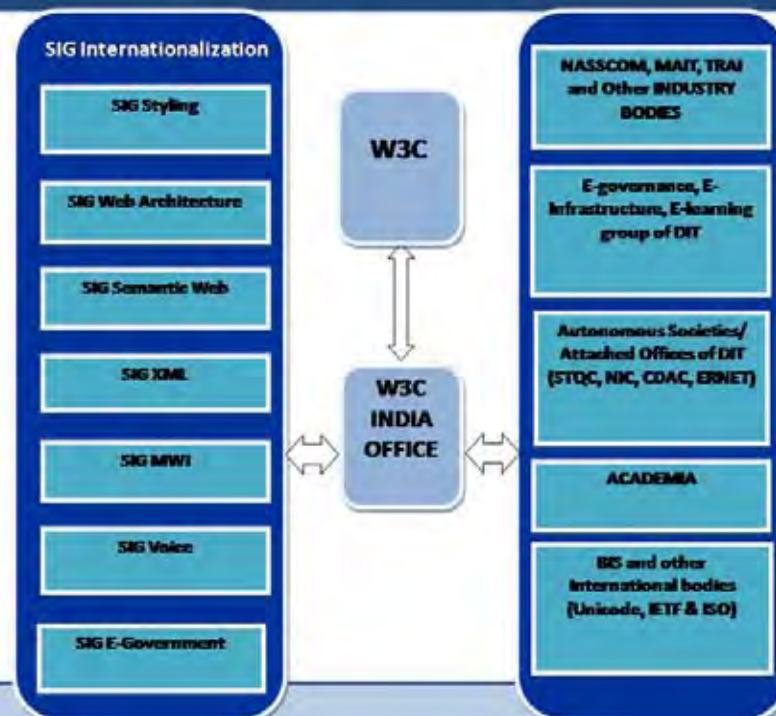


- To evolve vision for W3C India Office with the objective of maximizing the Indian participation in the standardization process of Web standards of W3C consortium with respect to 22 Constitutionally recognized Indian languages.
- To evolve Annual Road-map for W3C India Office.
- Evolving strategy for proliferation of W3C Standards in India through W3C India office.
- Guidance for organizing focused technical conferences in the areas of importance to ICT and Electronics Industry in India.
- Advise on setting up of Special Interest Groups in high priority areas.

9



## Methodology for implementation



10





## Work Done So Far



### WEB ACCESSIBILITY INITIATIVE (WAI) IN INDIA

- **National Informatics centre(NIC), Govt. of India**
  - formulated “Guidelines for Indian Government websites” in order to be internationally accepted standards to ensure that website are secure, user friendly and Universally accepted .
  - These guidelines is based on international standards including W3C's WCAG 2.0 and ISO 23026.

[See Here](#)
- **Centre for Internet and Society –** developing authorized translation of WCAG 2.0 Guidelines
- **STQC ,Organization of DIT , Govt. of India**
  - Implementing WCAG 2.0 Accessibility through Website Quality Certification

[See here](#)

11



## Mobile Web in India



### Issues for enabling Mobile web in India

- Character encoding
- Bandwidth and Cost
- Presentation Issues
- Input
- Issues at mobile device level
- Lack of standardization
- Fonts

12





## Mobile Web in India



### Messaging Issues

- Lack of availability for all characters.
- There is no guarantee that a message encoded will be displayed properly at the receiving terminal.
- Issue of Multiple Script -one language not addressed.
- Standardization of glyph support, syllable composition logic is also an important aspect and is dependent on the implementation level of handset manufacturer.
- Legacy Systems

13



## PRONUNCIATION LEXICON SPECIFICATION



Initiation of study for PLS 1.0 with respect to requirements for 2 Indian languages : Hindi ,Bangla and Tamil

### Issue regarding pronunciation of Hindi Language

- Homograph:
  - Most languages have words with different meanings but the same spelling (and sometimes different pronunciations), called homographs. For example, in Hindi the word कलक (धतूरा) and the word कलक (सोना) have identical spellings but different meanings.
- Homophones:
  - Similar pronunciation but different meanings (and possibly different spellings), for instance "बली" (बलवान) and "बलि" (बलिदान).
  - Similar pronunciation , similar meaning with different spelling  
"हिंदी" and "हिन्दी"
- Multiple Orthography for Indian Languages eg. Chillu character for Malayalam

14



### Styling issues in Indian languages

- first character
- Drop letter
- Bullets and numbering
- Collation
- Horizontal & Vertical arrangements of character
- Underlining of character
- Formatting issues

15

### National e-Governance Plan (NeGP)

- To Make all Government services accessible to the common man in his locality, through common service delivery outlets and to ensures :
  - Efficiency,
  - Transparency
  - Reliability of such services

16





## E-Governance applications in India



1. Bhoomi  
(<http://www.revdept-01.kar.nic.in/Bhoomi/Importance.htm>)
2. e-Seva  
(<http://esevaonline.com>)
3. Lok Mitra  
(<http://himachal.nic.in/lokmitra.htm>)
4. Mahiti Shakti  
(<http://www.apdip.net/resources/case/in13/view>)
5. Automatic Vehicle Tracking System  
(<http://www.mit.gov.in/default.aspx?id=599>)
6. MCD online  
(<http://www.mcdonline.gov.in>)

17



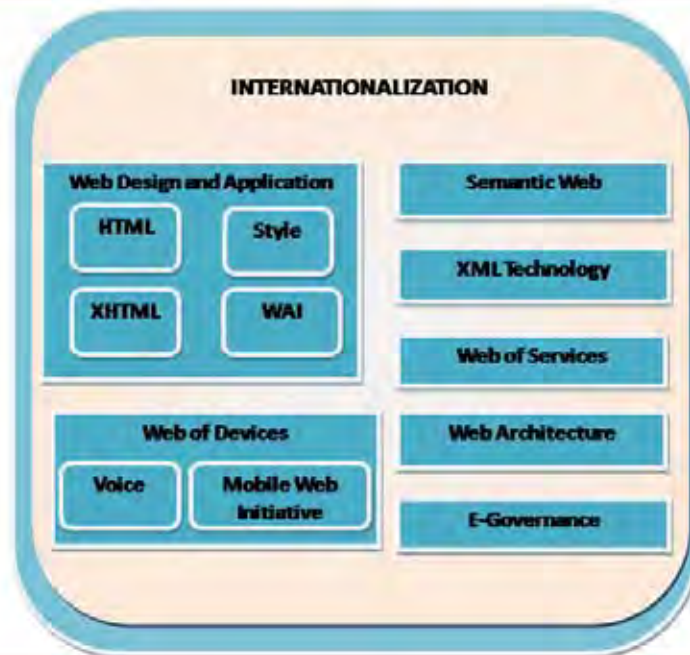
## Main Domains of W3C



- **Web design & applications**
  - CSS Level 2
  - Xhtml 1.1
  - HTML 5
  - WCAG 2.0
- **Web of Devices**
  - Mobile web Best practices
  - Pronunciation Lexicon specifications 1.0
  - SSML(Speech synthesis Markup Language) 1.0
- **Web Architecture & XML Technology**
  - XML 1.0
  - Extensible Stylesheet language (XSL) 1.1
- **Internationalization**
  - Internationalization Tag set 1.0
- **E-Government**
  - Use cases
- **Web of services**
  - Web Services Description language(WSDL) 2.0
  - Web services Addressing 1.0- Metadata
- **Semantic Web**
  - OWL Web ontology overview
  - Resource description framework (RDF)



- ❖ Internationalization
  - Internationalization Tag Set
- ❖ Web Design and Applications
  - Styling
  - Html
  - Xhtml
  - Wai
- ❖ Web of Devices
  - Mobile Web Initiative
  - Voice
- ❖ Semantic Web
  - OWL and RDF
- ❖ XML Technology
  - XML associated standards
- ❖ Web of Services
  - SOA
- ❖ Web Architecture
  - XML
- ❖ E-Government
  - Use cases



19



20





## Role Of W3C India Office



Technical feedback to W3C encompassing the usage of the Indian languages including internationalization issues and language specific feedback

Routing W3C standards with in Govt. and the Industry, Academia & provide feedback to W3C

Build a network of stake-holders (Industry, Government, Academia, public sector and media)

Having more participation & membership in India

23



Thank You

24





## Some Steps from the Web to a Semantic Web

Presented on

World Wide Web:

Technology, Standards and Internationalization 2010 Conference

by Klaus Birkenbihl, Coordinator World Offices, W3C

New Delhi, May 7<sup>th</sup> 2010

based on a slide set by Ivan Herman,  
Semantic Web Activity Lead, W3C

### Questions to answer

(2)

What is and why would we want a Web of data?

Why do we need standards for this?

-----

Can we use established WWW technology to do it?

What else would it take to build it?

Is it real?



**Let's e.g. organize a trip to  
Budapest using the Web!**

Copyright © 2010, W3C

(3)



**Let's find a proper flight ...**

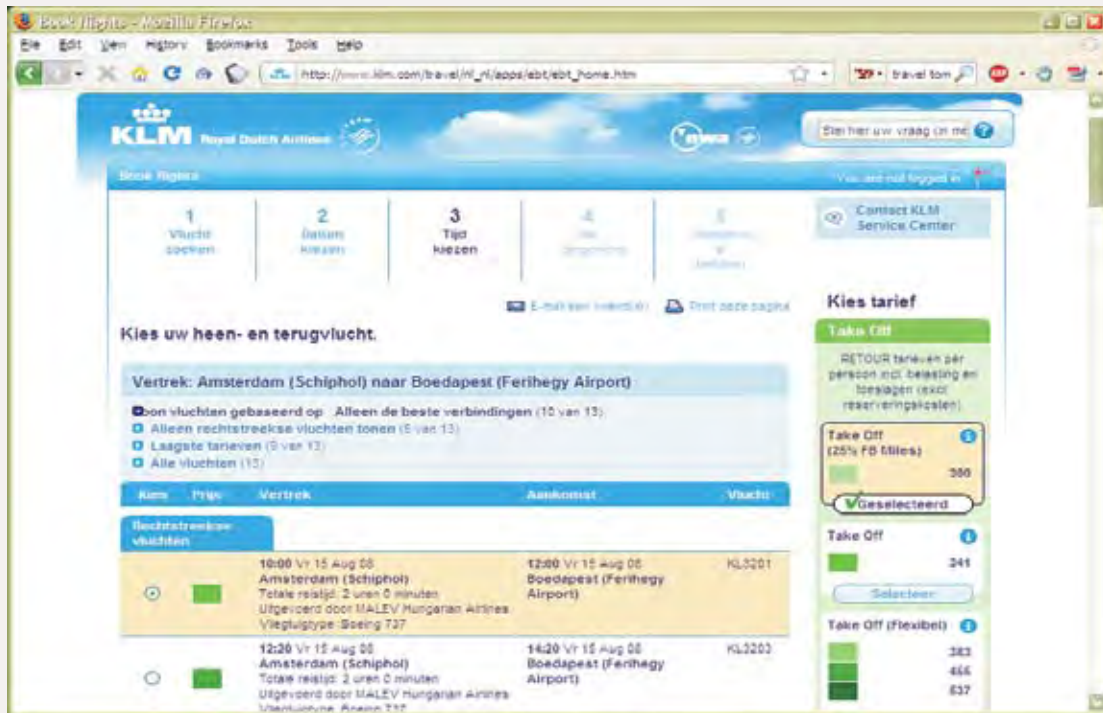
Copyright © 2010, W3C

(4)





... a big, reputable airline, or ...

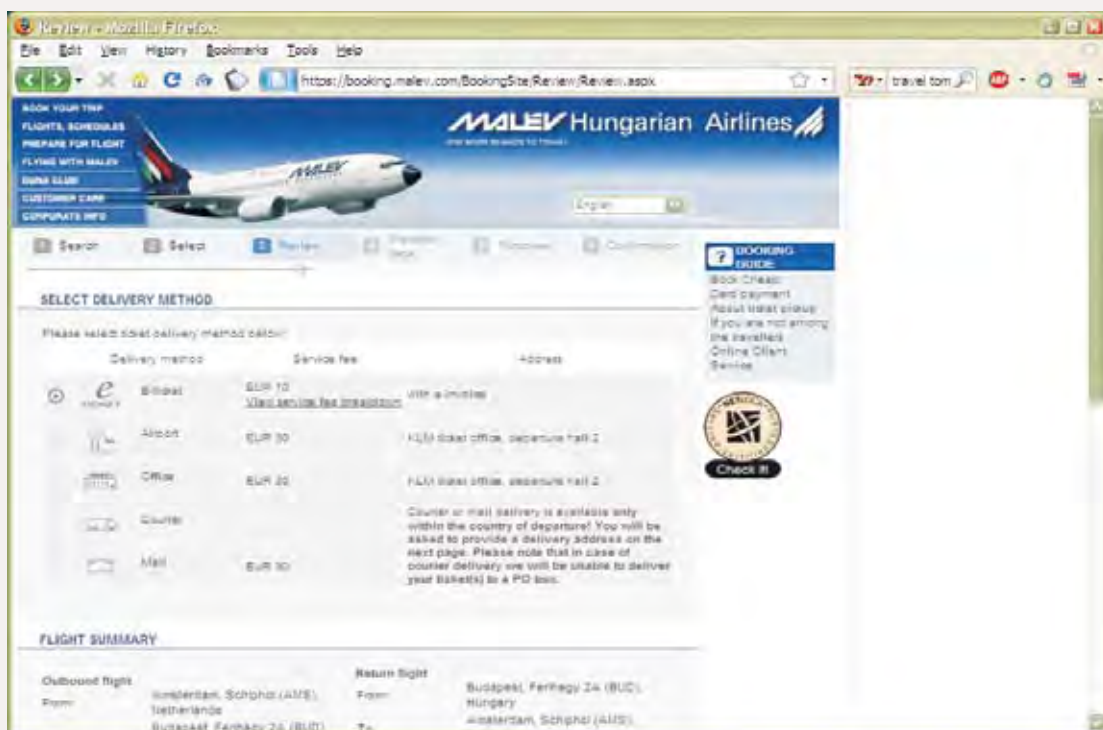


Copyright © 2010, W3C

(5)

W3C Semantic Web

... the airline of the target country, or ...



Copyright © 2010, W3C

(6)

W3C Semantic Web



... or a low cost one

The screenshot shows the Wizzair website interface. The header includes the Wizzair logo and a navigation menu with links like 'online booking', 'useful information', 'destinations', 'travel services', and 'partners'. A sidebar on the left contains links for 'flights', 'agency login', 'my account', 'search bookings', and 'log in'. The main content area displays flight search results for the route 'Eindhoven -> Budapest-Terminal 1'. A table lists available flights with columns for date, fareclass, flight number, departure, arrival, price including tax, and taxes and charges.

date	fareclass	flight	departs	arrives	price including tax	taxes and charges
Fri 18 Aug 08	Web	W6 228	13:25	15:20	Adult 94.99 EUR	26.00 EUR
Sat 19 Aug 08	Web	W6 228	13:25	15:20	Adult 73.99 EUR	26.00 EUR

Copyright © 2010, W3C

(7)

W3C Semantic Web

You have to find a hotel,  
so you look for...

Copyright © 2010, W3C

(8)

W3C Semantic Web



... or a really luxurious one, or ...

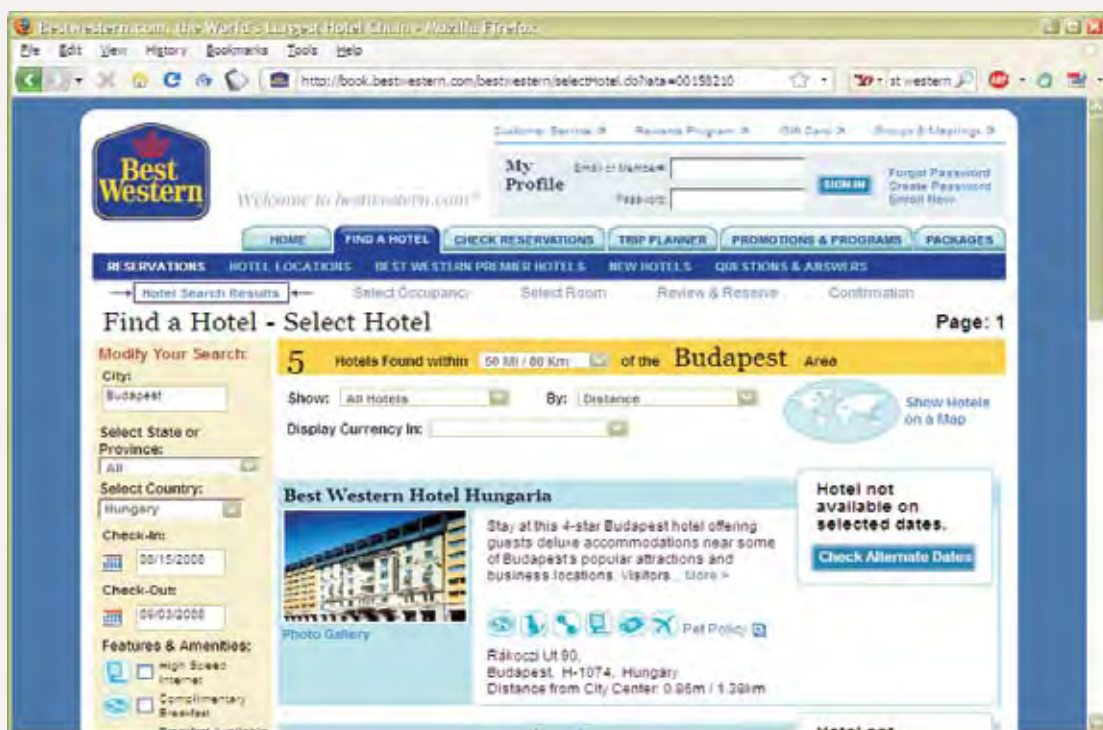


Copyright © 2010, W3C

(9)

W3C Semantic Web

... this one could work



Copyright © 2010, W3C

(10)

W3C Semantic Web



Of course,  
you could decide  
to trust a specialized site...

Copyright © 2010, W3C

(11)



... like this one, or...

(12)



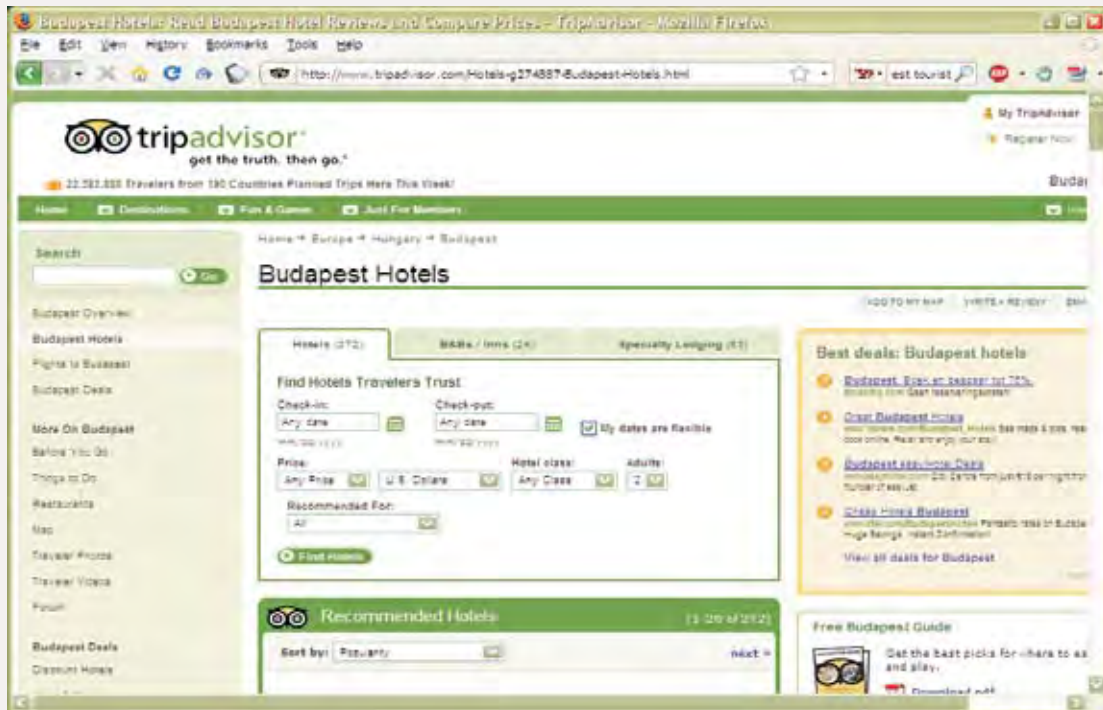
Copyright © 2010, W3C

(12)





... or this one



Copyright © 2010, W3C

(13)

W3C Semantic Web

you want to learn about Budapest



Copyright © 2010, W3C

(14)

W3C Semantic Web



## What happened here?

- You had to consult a large number of sites, all different in style, purpose, possibly language...
- You had to mentally *integrate* all this information to achieve your goals
- We all know that, sometimes, this is a long and tedious process!

Copyright © 2010, W3C

(15)



- All those pages are only tips of respective icebergs:
  - the real *data* is hidden somewhere in databases, XML files, spread sheets, SQL ...
  - you have only access to what the Web page designers allow you to see
- Specialized sites (Expedia, TripAdvisor) do a bit more:
  - they gather and combine data from other sources (usually with the approval of the data owners)
  - but they still control how you see those sources
- Sometimes you want more: you may want access to the original data and combine it yourself without remembering all intermediate results yourself!

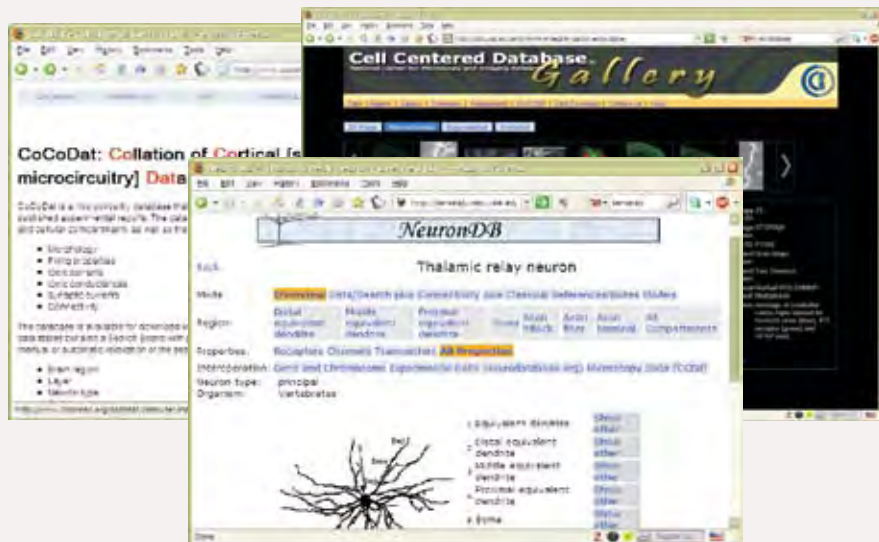
Copyright © 2010, W3C

(16)





## Same problem in research ...



- Companies may have to hire a person to answer questions based on those (public!) databases!

Copyright © 2010, W3C

(17)



## What would we like to have?

- We would like to have applications that can combine all the data in the different Web sites (or underlying databases) in a useful way.
- This would require that the applications can access the data
- This would require that the data can be linked like Web pages today

**Or put in another way:**

- We would like to *extend* the current Web with a "Web of data":
  - allow for applications to exploit the data directly

Copyright © 2010, W3C

(18)





But wait! Isn't this  
what mash-up sites are  
already doing?

Copyright © 2010, W3C

(19)



## Example: Managing trips (tripit.com)



Copyright © 2010, W3C

(20)





## How does it work

(21)

- Klaus forwards to Tripit the documents (mails, URIs) with the data related to a trip. e.g.
  - Flight bookings
  - Hotel reservations
  - Meetings
- Any time he has new documents he may add them
- Tripit tries to extract the data from these documents
- Based on the data it associates the documents to a trip
- It adds information from other sites about weather, directions, travel guides ...
- It checks its own database for travel activities of friends
- It connects with social networking sites
- It compiles a structured itinerary

Copyright © 2010, W3C

(21)



## This is great ... but

(22)


- Sometimes Tripit sends Klaus a message “Problem with your TripIt submission” and does nothing though the data was delivered
- Sometimes some data from in document are not identified
- Sometimes Klaus reads: “Please help us to improve! Let us know how good we captured your flight.”
- This gives a hint on what Tripit does: in case it does not know how to find the data it guesses
- Because there is no standardized way to access the data Tripit has to use proprietary interfaces and follow the changes – for all the many sources . *Greetings from Sisyphos to the programmers!*

Copyright © 2010, W3C

(22)





- So in some ways, mash-ups show the huge power of what a Web of data provides
- But mash-up sites are forced to do very ad-hoc jobs
  - various data sources expose their data via Web Services
  - each with a different API, a different logic, different structure
  - these sites are forced to reinvent the wheel many times because they don't use a standard way of doing things 

### Put it another way (again)...

- We would like to *extend* to the current Web with a **standardized** “Web of data”

Copyright © 2010, W3C

(23)



### Questions to answer

What is and why would we want a Web of data?

Why do we need standards for this?

-----

Can we use established WWW technology to do it?

URLs and Links

What else would it take to build it?

Machine readable description of Data (Metadata, Ontologies, Classification ...)

Is it real?

yes

Copyright © 2010, W3C

(24)





(25)

The following slides are not shown in this presentation (for the sake of time) but they detail the answers to the last 3 questions.

Slides at:

<http://www.w3.org/2010/Talks/0507NewDelhi-KB-IH/>

Copyright © 2010, W3C

(25)



(26)

## But what does this mean?

- What makes the current (document) Web work?
  - people create different documents
  - they give a globally unique address to it (i.e. a URI) and make it accessible to others on the Web

Copyright © 2010, W3C

(26)





## Then some magic happens... (27)

- Others discover the site and they link to it
  - So Search engines can find it and index it
- The more they link to it, the more important and well known the page becomes
  - remember, this is one criterion, search engines use to rank pages.
- This is the “Network effect”: some pages become important, and others begin to rely on it (even if the author did not expect it...)

Copyright © 2010, W3C

(27)



## Can this be used for a Web of Data? (28)

- Lessons learned: we should be able to:
  - “publish” the data to make it known on the Web
    - standard ways should be used instead of ad-hoc approaches
    - the analogous approach to documents: *give URI-s to the data*
  - make it possible to “link” to that URI from *other* sources of data (not only Web pages) using standard approaches
    - i.e., applications should not be forced to make targeted developments to access the data (as we saw with mash-ups)
    - generic, standard approaches should suffice
  - and let the network effect work its way...

Copyright © 2010, W3C

(28)





## But it is a little bit more complicated 😞

(29)

- On the traditional Web, humans are implicitly taken into account
- A Web link has a “context” that a person may use e.g. if you read on a Web page please [mail to Klaus Birkenbihl](#) ... you can guess that the link labelled “mail to Klaus Birkenbihl” leads you to his e-mail address.
- It all only works in a meaningful way if you, the human, can make correct and meaningful assumptions about the link.

Copyright © 2010, W3C

(29)



## Machines cannot interpret labels ...

(30)

- Something is missing in our model for the web of data!
- extra information (“label”) must be added to a link e.g. saying “this links to a Klaus Birkenbihl thing”.
- this information should be machine readable
- this label is a characterization (or “classification”) of *both* the link *and* its target
- in some cases, the classification should allow for some limited “reasoning”

Copyright © 2010, W3C

(30)





## Let us put together what we need for a Web of Data

- URI-s to publish data, not only full documents
- data can link to other data
- the data and the links (the “terms”) should be characterized/classified to convey some extra meaning
- standards for all these to maintain interoperability

Copyright © 2010, W3C

(31)



## So What *is* the Semantic Web?

Copyright © 2010, W3C

(32)





(33)

**It is a collection of standard technologies  
to realize a Web of Data**

Copyright © 2010, W3C

(33)



(34)

**It is that simple...  
but of course, the devil is in the details**

- a common model has to be provided for machines to understand the “labels” and draw some conclusions from that info
- the “classification” of the terms can become very complex for specific knowledge areas: this is where ontologies, thesauri, vocabularies, etc, enter the game...
- W3C has developed a set of standards for this
  - RDF – the Resource Description Framework
  - OWL – the Web Ontology Language (based on RDF)
  - SPARQL – a Query language for the Semantic Web
  - and a few more that make it easier to use

Copyright © 2010, W3C

(34)





(35)



(36)





(37)

All this sounds nice, but isn't that just a dream?

Copyright © 2010, W3C

(37)



(38)

## The “corporate” landscape is moving

- Major companies offer (or will offer) Semantic Web tools or systems using Semantic Web: Adobe, Oracle, IBM, HP, Software AG, GE, Northrop Gruman, Altova, Microsoft, Dow Jones, Google, Yahoo, Facebook, ...
- Others are using it (or consider using it) as part of their own operations: Novartis, Pfizer, Telefónica, ...
- Some of the names of active participants in W3C SW related groups: ILOG, HP, Agfa, SRI International, Fair Isaac Corp., Oracle, Boeing, IBM, Chevron, Siemens, Nokia, Pfizer, Sun, Eli Lilly, ...

Copyright © 2010, W3C

(38)





## Lots of Tools (not an exhaustive list!)

- Categories:
  - Triple Stores
  - Inference engines
  - Converters
  - Search engines
  - Middleware
  - CMS
  - Semantic Web browsers
  - Development environments
  - Semantic Wikis
  - ...
- Some names:
  - Jena, AllegroGraph, Mulgara, Sesame, flickurl, ...
  - TopBraid Suite, Virtuoso environment, Falcon, Drupal 7, Redland, Pellet, ...
  - Disco, Oracle 11g, RacerPro, IODT, Ontobroker, OWLIM, Tallis Platform, ...
  - RDF Gateway, RDFLib, Open Anzo, DartGrid, Zitgist, Ontotext, Protégé, ...
  - Thetus publisher, SemanticWorks, SWI-Prolog, RDFStore...
  - ...

Copyright © 2010, W3C

(39)



## In the end ...

- There is a huge potential for useful applications once we have a web of data
- Mash-ups give good examples for applications based on linking data
- When (re)designing a web site some thoughts on publishing data along with the documents might pay in the future.

Copyright © 2010, W3C

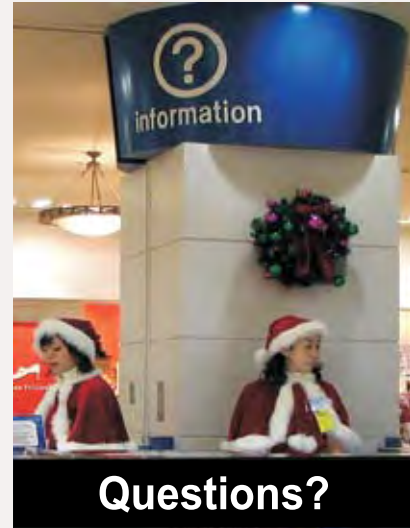
(40)







Thank you for your attention!



Questions?

- Slides are available at:  
<http://www.w3.org/2008/Talks/1002RioDeJaneiro-KB-IH/>  
in OpenDocument Presentation Format and PDF.





# Envisioning E-Governance with W3C Standards

By :  
Swaran Lata  
Country Manager, W3C India Office  
6, CGO Complex, Electronics Niketan  
New Delhi  
E-mail : [slata@mit.gov.in](mailto:slata@mit.gov.in)

## Table of Contents

- ❖ Goals
- ❖ Modalities for government on the web
- ❖ Limitations
- ❖ Examples
- ❖ The sample e-gov data with specific reference to localization
- ❖ Role of I18N Guidelines
- ❖ “Localization and Adoption of Open Standards”
- ❖ Adoption of Open Standards



- ❖ How Standards help in achieving target
- ❖ Unicode
- ❖ Common Locale Data Repository
- ❖ WAI – Web Accessibility Initiative
- ❖ CSS
- ❖ XHTML
- ❖ XML
- ❖ Accessing E-gov services through Mobile devices
- ❖ Semantic Web
- ❖ Conclusion

## GOALS



- ❖ Wider access of e-government services with localization support
- ❖ How I18N guidelines can help in achieving targets
- ❖ Enhance adoption of W3C (World Wide Web Consortium) standards to reach all sections of society.





## Modalities for Government on the Web

- ❖ **Provide:** public services on the web, either transactional or information services or both.
- ❖ **Engage:** with citizens and businesses, on government terms or on the citizens terms.
- ❖ **Enable:** public sector information re-use .

## Limitations



- ❖ The localization will become inhibiting for the applications which are not internationalized at design and development stage
- ❖ Due to limited Language Support / Lack of language Support , applications will not be accessible to the full satisfaction to non-English literate population
- ❖ Non conformity with Global Standards may result in fragmented access



## Broad categories of current e-gov applications

- ❖ Applications completely in English.
- ❖ Applications having static content in local language but dynamic content in English.
- ❖ Applications with complete local language support
- ❖ Multi-lingual application but only in limited languages (e.g., English and only one local language)
- ❖ Applications successfully running in one or two states and their mass scale replication
- ❖ Complimentary applications by private sector such as **e-choupal by ITC** and **knowledge network of NASSCOM foundation**

## Example e-gov application where registration forms are only in English



The screenshot shows the 'Online Submission of Applications' form. It includes fields for Applicant's Name, Sex (Male/Female), House No., Street/Sector/P.O. No., and Village/Locality. There is a dropdown menu for District (Central, New Delhi, Delhi). Below these fields, there is a section titled 'Click On Certificate you wish to Apply ONLINE' with a grid of buttons for various certificates: Registration of Marriage Certificate, Schedule Caste (SC) Certificate, Other Backward Classes (OBC) Certificate, Surviving Member Certificate, Certificate of Delhi Certificate, Order For Birth Certificate, Order For Death Certificate, Handicap Certificate, Income Certificate, Nationality Certificate, and Solvents Certificate. A 'Click Here To Apply Online' button is at the bottom.



## Static Content in Hindi but forms are in English



PERSONAL PARTICULARS FORM  
(In Duplicate)

Paste your  
cross signed  
recent colour  
photograph  
size 3.5"x5.5 cm

1. Full name (initials not allowed) \_\_\_\_\_

2. Sex: Male / Female / Others \_\_\_\_\_

3. Has the applicant ever changed name? \_\_\_\_\_

4. If yes, previous name: \_\_\_\_\_

4. Date of Birth: \_\_\_\_\_ 5. Place of Birth: \_\_\_\_\_

6. Profession: \_\_\_\_\_

7. a) Father: \_\_\_\_\_ (Surname) (Name)

b) Mother: \_\_\_\_\_

## English and local language application





## The sample e-gov data with specific reference to localization

Application	Monolingual/Bilingual/Multilingual
Karnataka Bhoomi	Bilingual ( English and Kannada)
India Portal	Bilingual (English and Hindi)
National Population Register	Multilingual (English, Kannada, Tamil, Gujarati )
Common Wealth Games Website	Monolingual (English)
Pay roll	Monolingual (English)
Financial applications (VAT, Accounts, e tendering )	Monolingual (English)

## The sample test data for localized applications under implementation

- ❖ In many applications local language implementation is only around 50%.
- ❖ Many fields that are in local language are transliteration of English words (e.g, invoice written in Kannada, Malayalam etc.) and not translations.

Continued



## Role of I18N in E-Gov

I18N guidelines are key for following requirements

- ❖ Browser Independent Application
- ❖ Easily Localizable Applications
- ❖ Improved Web search
- ❖ Effectively access Web content

### General architectural approach

- ❖ Will develop a set of XML, HTML and CSS source files referred to as repositories

### Structure of the techniques repository

- ❖ XSLT will be used to transform source files to XHTML in UTF-8 encoding



## “Localization and Adoption of Open Standards”

Local Language Interface is *“Not a desirable but an essential Component”*

- ❖ Language is the primary vector for communicating knowledge.
- ❖ Considering the multilingual and multi-script diversity in India thus it is imperative that, e-Governance applications need to be implemented with language framework.



- ❖ Need to adopt and follow standards on the language technology components for successful localization and wider access of Information and knowledge
- ❖ Adoption of World Wide Web Consortium (W3C) standard in respect of Localization
- ❖ Web based applications need to be developed in such a way that applications should be interoperable for seamless access of knowledge. W3C develops such technologies with specifications, guidelines, software and tools

Continued

- ❖ In order to enable multi-locale operation of the Web services and to create the ability for locale negotiation, this specification describes a standardized method for identifying locales and locale and/or language tags on the web
- ❖ Indian Language Web-Browsers need to be W3C complaint

Continued



## Adoption of Open Standards

**“Ability to exchange information and mutually use the information which has been exchanged.”**

- ❖ All the standards of W3C for multilingual and multi-modal interfaces such as HTML, XHTML, XML, CSS, and VOICE XML 2.0 are open standards.
- ❖ W3C offers a host of validation services using these standards.
- ❖ Adopting these open standards and localization process following W3C guideline may easily ensure interoperability between storage, display and access to service even in the multilingual paradigm.

## How Standards help in achieving target



<http://www.w3.org/>





## Unicode

- ❖ Seamless data storage and search if data is stored in UNICODE
- ❖ All 22 Officially recognized Indian Languages including Vedic Sanskrit represented in UNICODE
- ❖ **Declared as Text Encoding Standard for All E-Governance Applications**
- ❖ Most of the e-gov applications are intranet/internet based, hence, the need to migrate to Unicode is necessary, so that data can be seamlessly ported and accessed across the platforms which is not possible using ISCII
- ❖ The path W3C follows to making text on the Web truly global is Unicode



## Common Locale Data Repository

- CLDR is by far the largest and most extensive standard repository of locale data.
- **Provide locale data in the XML format for use in many e-gov applications**
- Locale Data for Indian Languages are in the process of modification
- Six Languages in CLDR **Hindi , Nepali, Bengali , Assamese, Malayalam and Gujarati** are finalized and uploaded to **UNICODE CLDR**
- Other languages in process.

```
<localeDisplayNames>
<language type="aa">अफार</language>
<language type="aa" alt="proposed-x1001"
draft="unconfirmed">अफार</language>
<language type="ab">अब्खाज़ियन्</language>
<language type="ach">अकोली</language>
<language type="ae"
draft="contributed">अवेस्तन</language>
<language type="af">अफ्रीकी</language>
<language type="afa">अफ्रो-एशियाई भाषाएँ</language>
<language type="afa" alt="proposed-x1001"
draft="unconfirmed">अफ्रो-एशियाई भाषाएँ</language>
<language type="afh">अफ्रीलीयाई</language>
```





## WAI – Web Accessibility Initiative

- ❖ An example of a successful education and outreach program that helps governments achieve compliance goals.
- ❖ WCAG 2.0 has 12 guidelines that are organized under 4 principles:
  - perceivable,
  - operable,
  - understandable, and
  - robust.
- ❖ **NIC has developed guidelines with the aim to assist the IT Managers of Government Departments in managing their websites in an effective and efficient manner.**



## WAI – Web Accessibility Initiative

### SUGGESTIONS

- ❖ Use of Unicode and Open Office fonts for Indian Languages
- ❖ Different color Selector for the people suffering from Color Blindness
- ❖ Use W3C Slide maker Tool to read and listen content slowly by user preference for easier understanding of the disable people
- ❖ Conversion Tools should be used like Text to Speech and Speech to Text for Blind and physically challenged people

Continued





## CSS - Cascading Style Sheet

- ❖ It's important to design a website for a wide range of audiences. This includes people with
  - disabilities,
  - people using mobile devices,
  - people with outdated technology.
- ❖ CSS standards *will avoid many of the accessibility issues*

### ISSUES IN INDIAN LANGUAGES

#### Bullets and Numbers

- ❖ Number schemes/ bulleting needs to be supported in Indian languages as well.



## CSS - Cascading Style Sheet

### Underlining of the characters

There is some examples of Indian languages in which Matra's are not readable due to underlining of characters.

- ❖ Hindi -
- ❖ Punjabi - \_\_\_\_
- ❖ Bengali - \_\_\_\_\_
- ❖ Gujarati - \_\_\_\_\_
- ❖ Marathi- \_\_\_\_\_

Continued





## CSS - Cascading Style Sheet

### Suggestions

- ❖ Implementation of CSS standards developed by W3C regarding Indian languages
- ❖ Standards however need to be provided to those developing CSS so that user could have the facility to use bulleting in his own Indic languages.

Continued



## XHTML

- ❖ XHTML was developed to make HTML more extensible and increase interoperability with other data formats.
- ❖ Websites should be designed and operated in accordance with the needs of users.
- ❖ Represent an important shift in Web applications, replacing complicated, hard-to-maintain scripting with declarative markup, and fully separating content from presentation.









## XML

### ❖ E-Filing

- revenue documents,
- court documents etc

### ❖ Law enforcement reporting

- arrest,
- transfer or
- release reports etc

### ❖ Licenses and permits

- Vehicle registration
- Driver's licenses
- Professional licenses

Continued



## Accessing e-gov services through mobile devices

*"Making Web access from a mobile device as simple as Web access from a desktop device."*

### Issues in Indian Languages

#### ❖ Messaging Issues

- Lack of availability for all characters.
- Issue of Multiple Script

#### ❖ Device Limitations

- Mobile browsers often do not support scripting or plug-ins, which means that the range of content that they support is limited.





## Accessing e-gov services through mobile devices

### ❖ Mobile Keypads

- Multi-tap issues
- Dictionary Based issues
- Transliteration issues

### Suggestion

Standardization of mobile media also required to be addressed taking into consideration of specific requirements of each of Indic languages.

Continued



## Semantic Web Implementation in E-governance

- ❖ Improving access of resources.
- ❖ Enable people to create data stores on the Web, build vocabularies, and write rules for handling data.
- ❖ Linked data are empowered by technologies such as RDF, SPARQL, OWL, and SKOS.
- ❖ **E-Government agencies can share and exchange information through semantic web**



## Conclusion

- ❖ E-Governance can transform citizen services, provide access to information to empower citizens, enable their participation in government and enable access to economic and social opportunities
- ❖ The issues that primarily affect the e-governance applications are, interoperability between the heterogeneous systems and data repositories, absence of standard taxonomy, absence of compliance to best practices for seamless web-accessibility and lack of internationalization & localization of the multimodal application interface
- ❖ Implementation of W3C Standards will be a prime factor for seamless and interoperable solutions to achieve the goal for all 22 constitutionally recognised Indian languages



!! QUESTIONS !!



😊 THANK YOU 😊



## Design of Efficient Hindi Keypad for Mobile Hand-held Device

Devendra Jalihal  
dj@iitm.ac.in

Department of Electrical Engineering  
Indian Institute of Technology Madras  
Chennai 600 036

World Wide Web: Technology, Standards and Internationalisation  
Conference, 2010, New Delhi

### Acknowledgements:

Karthik Aditya, (IITM) UC Berkeley

Vivek Pani, Reverie

Nadeem Akhtar and Babu Narayanan, CEWiT/BWCI



## Outline

- ① Motivation
- ② Metrics for Comparison
- ③ Simulation
- ④ Summary

## Motivation

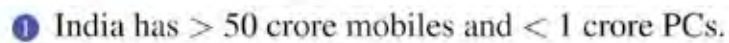
### Motivation

Mobile phone is the primary device of access

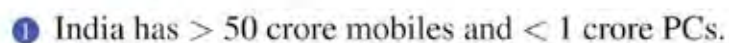




Mobile phone is the primary device of access



Mobile phone is the primary device of access



## ② Crystal gazing



Motivation

## Motivation

Mobile phone is the primary device of access



- ① India has > 50 crore mobiles and < 1 crore PCs.
- ② Crystal gazing
  - Mobile to be the most common web access device by 2013. Gartner, Jan 2010.
  - 3 billion people will transact electronically via mobile or Internet technology. Gartner, Jan 2010.
  - In India, Internet access will be Mobile > TV Remote > PC.



Motivation

## Motivation

Services in Offing

- ① Mobile Payment Gateway
  - driven by banks and other financial institutions
  - coordinated by RTBI/IITM
  - has defined protocols for financial transactions with and without pre-registered parties
- ② Farm Advisory Program (funded by NAIP at IITM)
  - 1200 farmers in three districts of TN use mobiles to interact with experts in a multi-way communication





Motivation

## Motivation

Problem of Language

### ① If language of access == English

- Keypad standard ITU E.161 maps 26 letters to 9 keys
- host of tools: dictionary support, readable fonts, predictive text, . . .
- 7-bit ASCII  $\Rightarrow$  160 character SMS  $\Rightarrow$  efficiency on Air



ITU E.161

Motivation

## Motivation

Problem of Language

### ① If language of access == English

- Keypad standard ITU E.161 maps 26 letters to 9 keys
- host of tools: dictionary support, readable fonts, predictive text, . . .
- 7-bit ASCII  $\Rightarrow$  160 character SMS  $\Rightarrow$  efficiency on Air



ITU E.161

### ② If language of access == Hindi or any Indian Language

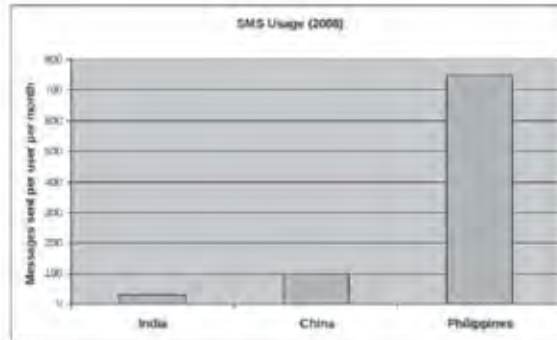
- no one *standard keymap*, no standard editing guidelines
- highly non-linear scripts  $\Rightarrow$  vendors resorted to English-centric font solutions
- minimal dictionary, prediction and transliteration support
- 7-bit code for Indic languages recently standardised in 3GPP



Motivation

## Motivation

Consequence: Low SMS Usage in India



Source:

Philippines & China: Mobile Messaging Futures 2009-2013, Portio Research

India: The Indian Telecom Services Performance Indicators October – December, 2008, TRAI

Motivation

## Motivation

Problem of Plenty

1. ०-९ 0-9	2. अ-क ABC	3. ए-ओ DEF
4. क-ड GHI	5. च-ज JKL	6. ट-ण MNO
7. त-न PQRS	8. प-म TUV	9. य-र WXYZ
* +	0	#

Nokia

1. ०-९ 0-9	2. अ-क ABC	3. ए-ओ DEF
4. क-ड GHI	5. च-ज JKL	6. ट-ण MNO
7. त-न PQRS	8. प-म TUV	9. य-र WXYZ
* +	0	#

LG

1. ०-९ 0-9	2. अ-आ ABC	3. क-ख DEF
4. च-छ GHI	5. इ-ए JKL	6. ट-ठ MNO
7. त-थ PQR	8. र-ल STU	9. द-ड VWX
* +	0	#

Sony Ericsson

1. ०-९ 0-9	2. अ-आ ABC	3. क-ख DEF
4. च-छ GHI	5. इ-ए JKL	6. ट-ठ MNO
7. त-थ PQR	8. र-ल STU	9. द-ड VWX
* +	0	#

Spice

Figure: Keymaps for Hindi from handset vendors



Motivation

## Motivation

Problem of Plenty

1 ० ○ ○	2 अ-क ABC	3 ए-ओ DEF
4 क-ड GHI	5 च-झ JKL	6 ट-ण MNO
7 त-न PQRS	8 प-म TUV	9 य-र WXYZ
* ○	0 ○	# ○

Nokia

1 ०	2 अ-क	3 ए-ओ
4 क-ड	5 च-झ	6 ट-ण
7 त-न	8 प-म	9 य-र
* +	0	#

LG

1 ०	2 अ-आ	3 क-इ
4 क-उई	5 इ-एउ	6 ट-फउ
7 ब-वए	8 र-नओ	9 य-इओ
* ○	0	#

Sony Ericsson

1 ०	2 अ-आ	3 क-इ
4 क-उई	5 इ-एउ	6 ट-फउ
7 ब-वए	8 र-नओ	9 य-इओ
* ○	0	#

Spice

Figure: Keymaps for Hindi from handset vendors

### Questions:

1. Is there an efficient mapping of IL alphabets to the 12-key device?
2. How does one measure, compare and evaluate various mappings?

1 ० ○ ○	2 अ-क ABC	3 ए-ओ DEF
4 क-ड GHI	5 च-झ JKL	6 ट-ण MNO
7 त-न PQRS	8 प-म TUV	9 य-र WXYZ
* ○	0 ○	# ○

Nokia

1 ०	2 अ-क	3 ए-ओ
4 क-ड	5 च-झ	6 ट-ण
7 त-न	8 प-म	9 य-र
* +	0	#

LG

1 ०	2 अ-आ	3 क-इ
4 क-उई	5 इ-एउ	6 ट-फउ
7 ब-वए	8 र-नओ	9 य-इओ
* ○	0	#

Sony Ericsson

1 ०	2 अ-आ	3 क-इ
4 क-उई	5 इ-एउ	6 ट-फउ
7 ब-वए	8 र-नओ	9 य-इओ
* ○	0	#

Spice

Figure: Keymaps for Hindi from handset vendors

### Questions:

1. Is there an **efficient** mapping of IL alphabets to the 12-key device?
2. How does one measure, compare and evaluate various mappings?
  - **efficient** = composing message is easy and fun



## Metrics for Comparison

## ① Keypad Clash Metric

- Indian languages map 6-7 characters per keys  $\Rightarrow$  high degree of multi-tapping
- *disambiguation/predictive texting* is needed
  - to succeed, number of words mapped to key combination be small
  - English : 4-6-6-3  $\Rightarrow$  *home, good, gone*
- A mapping with *smaller value for Keypad Clash aids prediction and is better*

## ② Keypad Distance Metric

- hand-held is a one-finger device  $\Rightarrow$  distance travelled is a measure of effort
  - 4-5-3-9  $\Rightarrow$  the distance travelled =  $1 + \sqrt{2} + 2 \approx 4.4$
- A mapping with *smaller value for Keypad Distance is better*

## ③ Keypad Learning Metric

- Language learning requires a *small number of words* to be learned in a *small number of attempts*
- A mapping that *helps the user learn the language faster* is better

## Metrics for Comparison

## ① Keypad Clash Metric

- Indian languages map 6-7 characters per keys  $\Rightarrow$  high degree of multi-tapping
- *disambiguation/predictive texting* is needed
  - to succeed, number of words mapped to key combination be small
  - English : 4-6-6-3  $\Rightarrow$  *home, good, gone*
- A mapping with *smaller value for Keypad Clash aids prediction and is better*

## ② Keypad Distance Metric

- hand-held is a one-finger device  $\Rightarrow$  distance travelled is a measure of effort
  - 4-5-3-9  $\Rightarrow$  the distance travelled =  $1 + \sqrt{2} + 2 \approx 4.4$
- A mapping with *smaller value for Keypad distance is better*

## ③ Keypad Learning Metric

- Language learning requires a *small number of words* to be learned in a *small number of attempts*
- A mapping that *helps the user learn the language faster* is better



## Metrics for Comparison

### 1 Keypad Clash Metric

- Indian languages map 6-7 characters per keys  $\Rightarrow$  high degree of multi-tapping
- disambiguation/predictive texting is needed
  - to succeed, number of words mapped to key combination be small
  - English : 4-6-6-3  $\Rightarrow$  home, good, gone
- A mapping with smaller value for Keypad Clash aids prediction and is better

### 2 Keypad Distance Metric

- hand-held is a one-finger device  $\Rightarrow$  distance travelled is a measure of effort
  - 4-5-3-9  $\Rightarrow$  the distance travelled =  $1 + \sqrt{2} + 2 \approx 4.4$
- A mapping with smaller value for Keypad distance is better

### 3 Keypad Learning Metric

- Language learning sequence: svaras, anusvara, visarga, vargiya vyanjanas, avargiya vyanjanas
- A mapping that follows this sequence is better, needs less learning



## Simulation

### Method

- Lexical list of 7500+ commonly used Hindi words
- Define three thresholds:  $T_C$ ,  $T_D$ ,  $T_L$ , for Clash, Distance, learning metrics
- Procedure:
  - $T_C \leq 3 \Rightarrow$  That mapping is Desirable else Undesirable.
  - $T_D \leq 8 \Rightarrow$  That mapping is D else U.
  - $T_L \leq 8 \Rightarrow$  That mapping is D else U.





## Simulation

Results - Clash Metric



Nokia



Samsung



Sony Ericsson

### 1 Observations

- Sony-Ericsson spreads vowels and scores on *clash metric*.
  - E.161 does it for English
- Hindi and other ILs have large number of nasal sounds
  - Examples: *mein*, *nahin* etc.
- If nasals are mapped to a separate key, *clash* will reduce.

## Simulations

Proposed keypad mapping

### Proposed Keymap





## Simulations

### Results

Make	$D$ in %
Nokia	69
Samsung	65
Sony Ericsson	93
Proposed	90

Table: *Keypad Clash Metric* evaluation as %  $D$  for  $T_C = 3$ .

Make	$D$ in %
Nokia	39
Samsung	41
Sony Ericsson	49
Proposed	55

Table: *Keypad Distance* evaluation as %  $D$  for  $T_D = 8$ .

## Results

Make	$D$ in %
Nokia	93
Samsung	85
Sony Ericsson	64
Proposed	88

Table: *Keypad Learn Metric* evaluation as %  $D$  for  $T_L = 8$ .



## Results

### 1 Observations

- A mapping which scores high on one metric performs poorly on other metric.
- Mathematically impossible to design a mapping that scores uniformly high on all metrics.
- Assign weights to each metric judiciously to make a *subjective judgement* on suitable mapping.

## Summary

### 1 Proposed keypad mapping suited for predictive-texting

- distributes vowels over the 9 keys
- maps consonants based on their *vargas*, and nasals separately
- maps the *avargiya* consonants in the most natural way
- also scores reasonably high on other metrics



## Summary

- ❶ Proposed keypad mapping suited for predictive-texting
  - distributes vowels over the 9 keys
  - maps consonants based on their *vargas*, and nasals separately
  - maps the *avargiya* consonants in the most natural way
  - also scores reasonably high on other metrics
- ❷ Indian Languages are phonetic.
  - Our simulations with Tamil, Marathi etc. show it works for other languages too

## Summary

- ❶ Proposed keypad mapping suited for predictive-texting
  - distributes vowels over the 9 keys
  - maps consonants based on their *vargas*, and nasals separately
  - maps the *avargiya* consonants in the most natural way
  - also scores reasonably high on other metrics
- ❷ Indian Languages are phonetic.
  - Our simulations with Tamil, Marathi etc. show it works for other languages too
- ❸ Many minor variations possible. BWCI/CEWiT is working on a recommendation.



## Summary

For increase in the use of Indian languages on mobile devices

- ① standardised keypad mappings
- ② 7-bit Indian Language encoding for efficiency
- ③ proper display fonts, standard editing procedures
- ④ dictionary, prediction and transliteration support, and
- ⑤ appropriate policy support

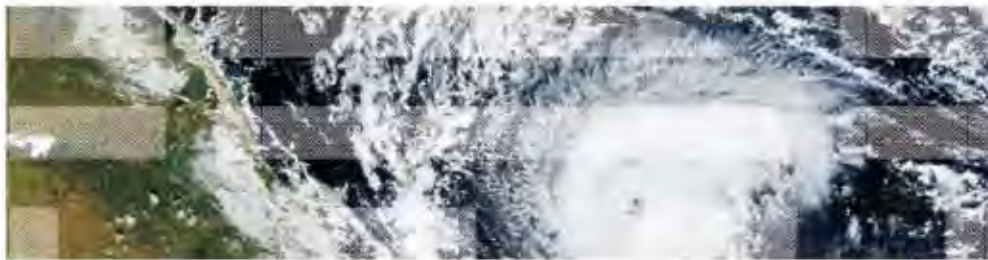


Kiran Kate, Karthik Visweswariah, **Nanda Kambhatla**

IBM Research - India, Bangalore



## Bridging the Language Divide using Machine Translation: Advances in English to Hindi Translation



© 2019 IBM Corporation



### Digital Divide and Language Divide

- Digital Divide
  - Between people with access to Internet and computing resources and those who do not
  - People getting disenfranchised
- Language Divide
  - Between any sets of people who do not speak a common language
  - Unable to interact with service providers from large sections of the population
- Machine Translation
  - Automated translation of documents, speech from one language to another
  - MT engines good enough to enable browsing of content expressed in other languages
  - Can help bridge Language Divide and Digital Divide





## Statistical Machine Translation

- English to Hindi translation: find a Hindi sentence  $h$  such that  $P(h|e)$  is maximized
- Learnt from a corpus of English-Hindi data
- Most likely translation:  $\text{argmax}_h P(e|h) P(h)$
- **Language Model** : probability of occurrence of Hindi sentence  $h$
- **Translation Model** : probability of  $e$  being expressed as Hindi sentence  $h$
- Search for Hindi sentence  $h$  that maximizes  $P(h) * P(e | h)$
- Parameters of models estimated from large database of  $(e, h)$  sentence pairs using a statistical algorithm

3

© 2010 IBM Corporation



## Language Model

$$\begin{aligned} \Pr(s_1 s_2 \dots s_n) \\ = \Pr(s_1) \Pr(s_2 | s_1) \dots \Pr(s_n | s_1 s_2 \dots s_{n-1}). \end{aligned}$$

- Compose probability of sentence  $S$  from probability of single words given all preceding words
- Too many histories of preceding words – too many parameters
- Simplifying assumption
  - Equivalence classes of histories
  - $N$ -gram model : histories equivalent if agreeing in final  $n-1$  words
  - Typically trigram models are used, 4-gram models are becoming tractable

4

© 2010 IBM Corporation





## Translation Model

- Probability that Hindi word  $h$  is translation of English word  $e$
- Compute  $P(h | e)$  from parallel corpus
- Statistical approaches rely on the co-occurrence of  $e$  and  $h$  in the parallel data
  - If  $e$  and  $h$  tend to co-occur in parallel sentence pairs, they are likely to be translations of one another

5

© 2019 IBM Corporation



## Unique challenges for Hindi (and other Indian languages)

- Lack of large parallel corpora
- Word order challenges
  - Large movements of words from source language order
- Linguistic resources not yet mature or ready
  - Parsing treebanks
  - Annotated POS corpora
  - Standardized guidelines for annotation

6

© 2019 IBM Corporation





## Extracting parallel corpora from comparable corpora

- For building language models, we require large corpora in target language
  - Relatively straightforward
- For building translation models, we require large amounts of bilingual parallel data
  - Document/sentence pairs which are translations of each other
  - More problematic
- Quality of SMT systems is dependent on amount of parallel data
- Comparison
  - German-English 1.3M sentence pairs
  - French-English 1.3M sentence pairs
  - Portuguese-English 1.2M sentence pairs
  - Hindi-English lagging far behind
- Extraction of sentence pairs from crawled 'comparable' data
  - IBM Model-1 trained on initial parallel data
  - Each candidate sentence assigned score using Model-1 probabilities
  - For each source sentence, target sentence with highest score is selected
  - Threshold applied to scores of selected sentence pairs

7

© 2010 IBM Corporation



## Supervised word alignments

- Early work on generating word alignments has been unsupervised
  - e.g. IBM Models 1-5, HMM models
- Recent work shows significant improvements using human annotated word alignments
- Maximum Entropy model
  - Features: lexical pairs, segmentation, WordNet, sounds-like, spelling, ...
  - Factors to keep alignments of adjacent words close together
  - Factors penalizing one word aligning to a large number of words of the other language

Aligner	Precision	Recall	fMeasure	BLEU
HMM	68.40	54.50	60.65	15.99
ME	84.20	71.16	77.14	17.47

8

© 2010 IBM Corporation





## Example

Hindi Sentence:

वे एक कुशल राजनेता के साथ – साथ एक अच्छे विद्वान , अर्थशास्त्री और विचारक भी हैं ।

Translated English Sentence: HMM Based System:

they , along with the best skilled scholar – politician , economist and thinker .

Translated English Sentence: HMM+ME Based System:

he 's a good politician as well as a good scholar , economist and thinker .

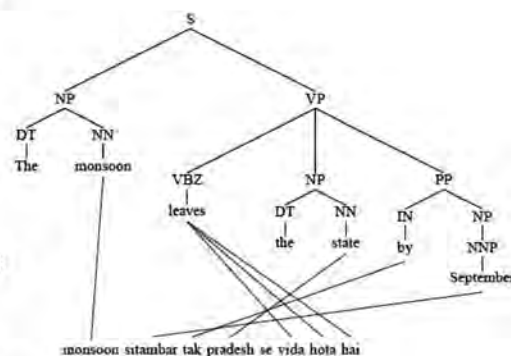
8

© 2010 IBM Corporation



## Syntax based re-ordering

- Different word order permutations in different languages is a fundamental challenge
- Hindi word order – Subject Object Verb
- English word order – Subject Verb Object
- Prepositions in English become postpositions in Hindi and appear after the noun phrase
- Words that are close in English can move arbitrarily far apart in Hindi



10

© 2010 IBM Corporation



IBM

## Syntax based re-ordering (contd.)

- Reordering model conditioned on source side parse tree
  - learned using parallel corpus of sentence pairs, machine generated word alignments, and source side parses
- Probabilistic model  $P(T|S)$ 
  - Word order in trees  $T$  assigned higher probability match the order in target language
  - Assume children of a node are ordered independently of all other nodes in the tree
  - Assume reordering at a node is dependent only on labels of its children
  - Re-ordering of children of a node modeled using log-linear formulation
    - simple count based statistical model with our choice of features.
- Source side sentences reordered by choosing  $\text{argmax}_T P(T|S)$ 
  - Reordering each interior node based on most frequent reordering of the constituents seen in training
- BLEU score with reordering is 21.7 as against BLEU score of 20.0 for the baseline system

11

© 2010 IBM Corporation

IBM

## Conclusions

- Challenges for Hindi-English Machine Translation
  - Lack of large parallel corpora
  - Lack of maturity of linguistic resources
  - Different word order issues
- Improved system but still low final score compared to other languages
  - Better approaches to leveraging comparable corpora
  - Utilize linguistic resources
- What will lead to next advances?
  - Creating/nurturing/extracting large bilingual parallel data
  - Additional linguistic resources like treebanks/probanks for Hindi

12

© 2010 IBM Corporation



**TATA** CONSULTANCY SERVICES



World Wide Web: Technology, Standards and Internationalization Conference, 2010  
W 3 C – India

## E-Gov use Cases Scenarios

**Tanmoy Chakrabarty**  
Vice-President  
&  
Global Head- Government ISU  
Tata Consultancy Services

New Delhi , 7 May, 2010

Experience, certainty,  
IT Services  
Business Solutions  
Outsourcing

### Citizens cannot travel long distances to avail services



**TATA** CONSULTANCY SERVICES

2

Information  
CONFIDENTIAL



## Online kiosks take services to citizens doorsteps



**MPOnline Limited**  
Authorized KIOSK List | Contact us | Feedback | Welcome Guest | Home | Login

Forms Download | GOs/Acts | Govt. Telephone / e-Mail | Madhya Pradesh Profile | KIOSK Registration Status | MPOnline Survey

**'Anytime Anywhere' services**

User ID:   
Password:   
Login Type:   
Login

**Single sign-on and single point access to multiple government services**

**Transparent governance leading to improved image of the government**

**NEW LAUNCHES**

- Scalable kiosks model under Public-Private Partnership mode providing employment to the local citizens
- It's citizens in terms of providing them with the facilities at their doorstep, in the process eliminating the need for coming to the Government offices and avoid the unnecessary
- Providing timely and comprehensive services to citizens in remote areas of the state

**MPOnline awarded the NASSCOM CNBC IT User award 2009**

Examination form without Late fee from 22nd Jan to 4th Feb 2010

मध्य प्रदेश - बहुमुखी विकास



## Life cycle of e-Services

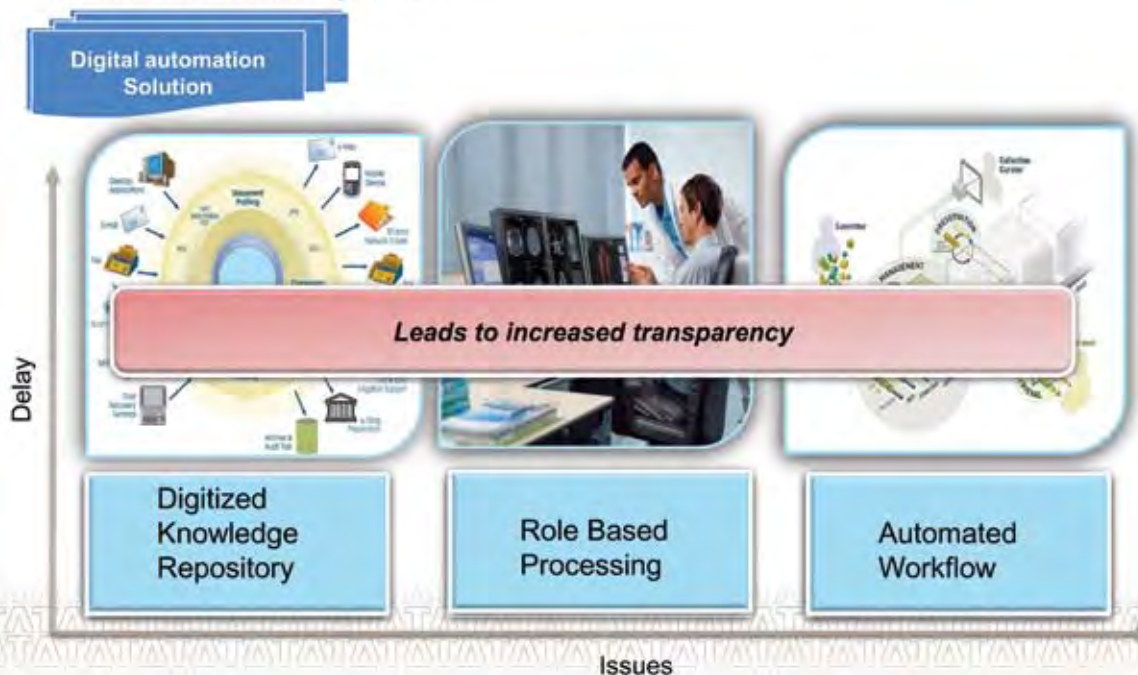


TATA CONSULTANCY SERVICES

5

TCS Confidential

## With a Digital Workflow automation at the back end to complete the service delivery chain



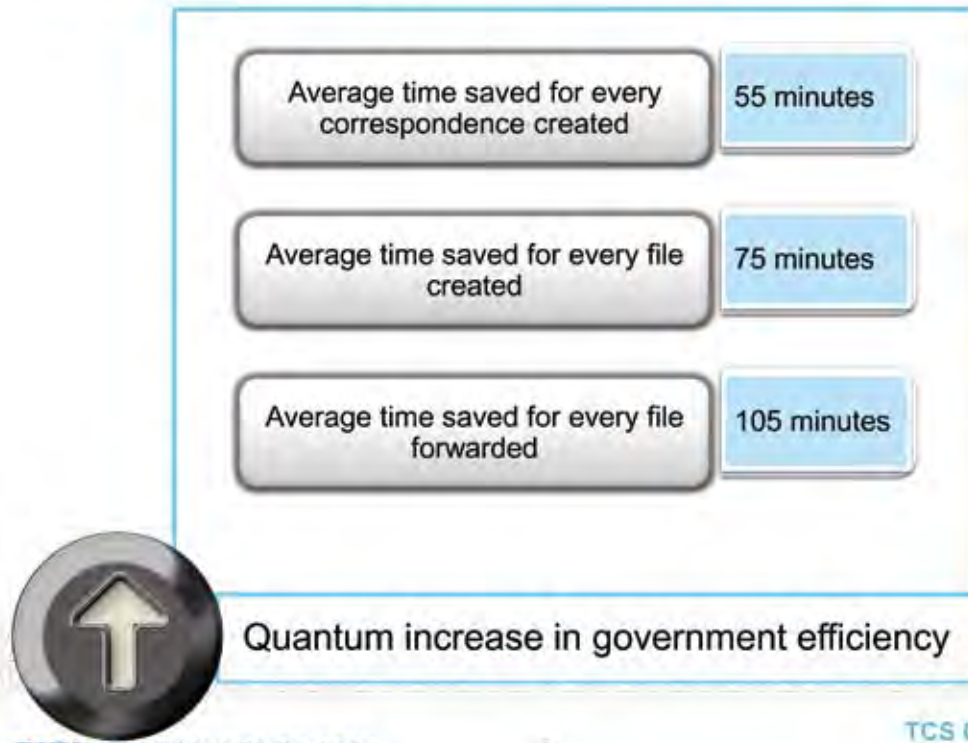
TATA CONSULTANCY SERVICES

6

TCS Confidential



## DigiGov – An Office Automation solution for Government



TATA CONSULTANCY SERVICES

TCS Confidential

## States with the highest GSDP growth rates!

State	Last 5 year growth rates**	DigiGov Users
Gujarat	11.05%	✓
Bihar	11.03%	✓

The Government of the top two States in India, with highest GSDP growth rate, have implemented Digital automation solutions for Government departments



TATA CONSULTANCY SERVICES

\*\*Source: <http://timesofindia.indiatimes.com/biz/india-business/Bihar-grew-by-1103-next-only-to-Gujarat/articleshow/5405973.cms>

TCS Confidential

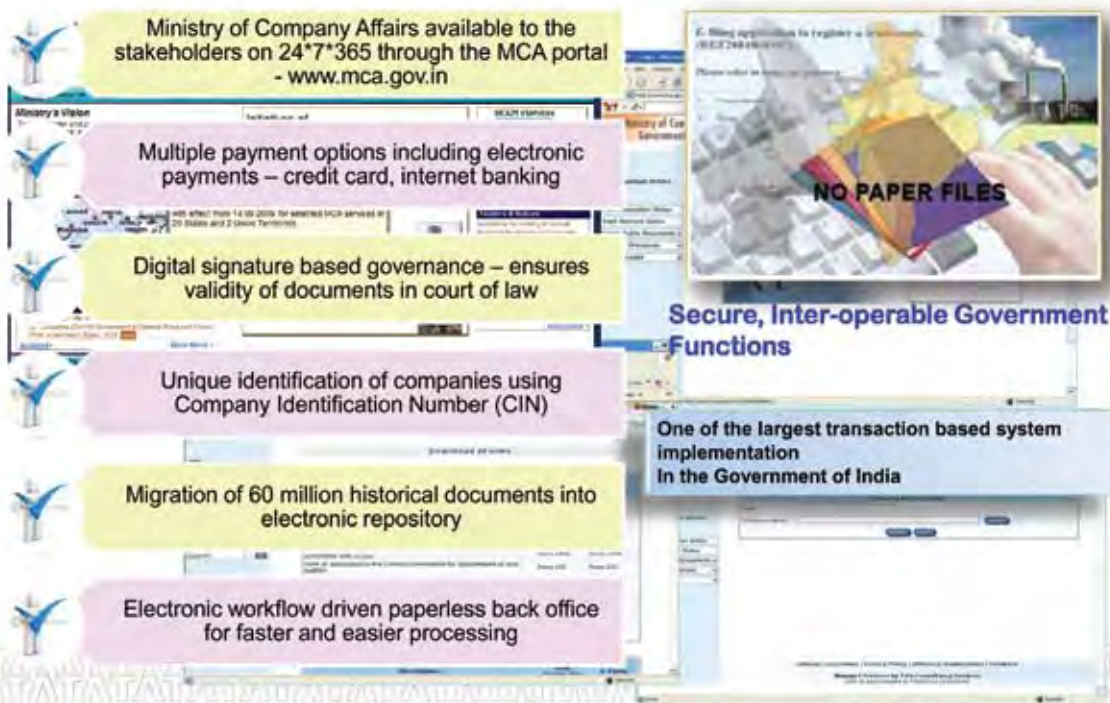


## Enabling Business through a Public Private Partnership Model

TATA CONSULTANCY SERVICES

TCS Confidential

### MCA21- Enabling the service transformation



Ministry of Company Affairs available to the stakeholders on 24\*7\*365 through the MCA portal - [www.mca.gov.in](http://www.mca.gov.in)

Multiple payment options including electronic payments – credit card, internet banking

Digital signature based governance – ensures validity of documents in court of law

Unique identification of companies using Company Identification Number (CIN)

Migration of 60 million historical documents into electronic repository

Electronic workflow driven paperless back office for faster and easier processing

**NO PAPER FILES**

**Secure, Inter-operable Government Functions**

**One of the largest transaction based system implementation in the Government of India**

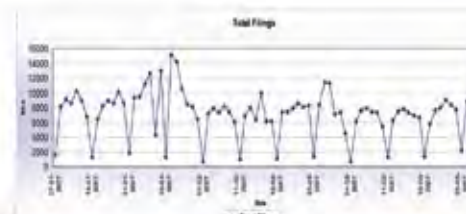
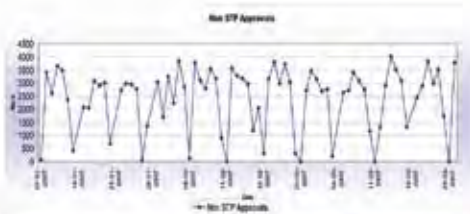
TATA CONSULTANCY SERVICES

- 10 -

TCS Confidential



## Some facts

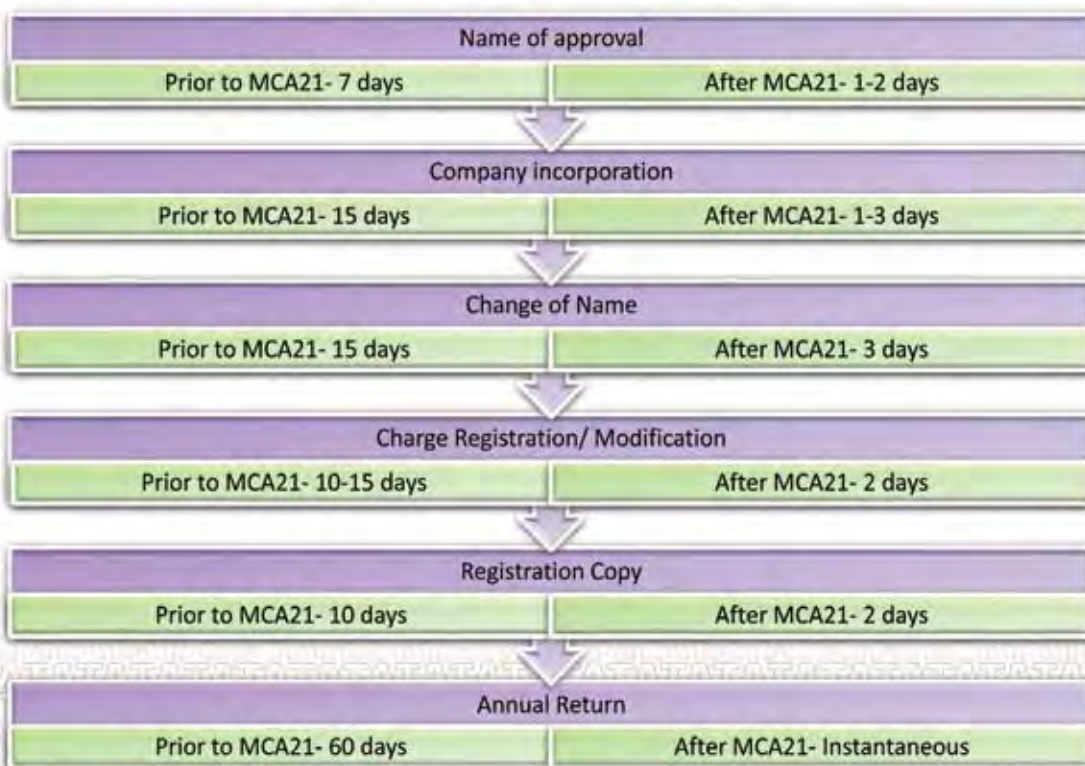


**Both e Filing and ROC approvals are happening at steady pace**

TATA CONSULTANCY SERVICES

11

## Increased efficiency in service delivery to customers



TATA CONSULTANCY SERVICES

12

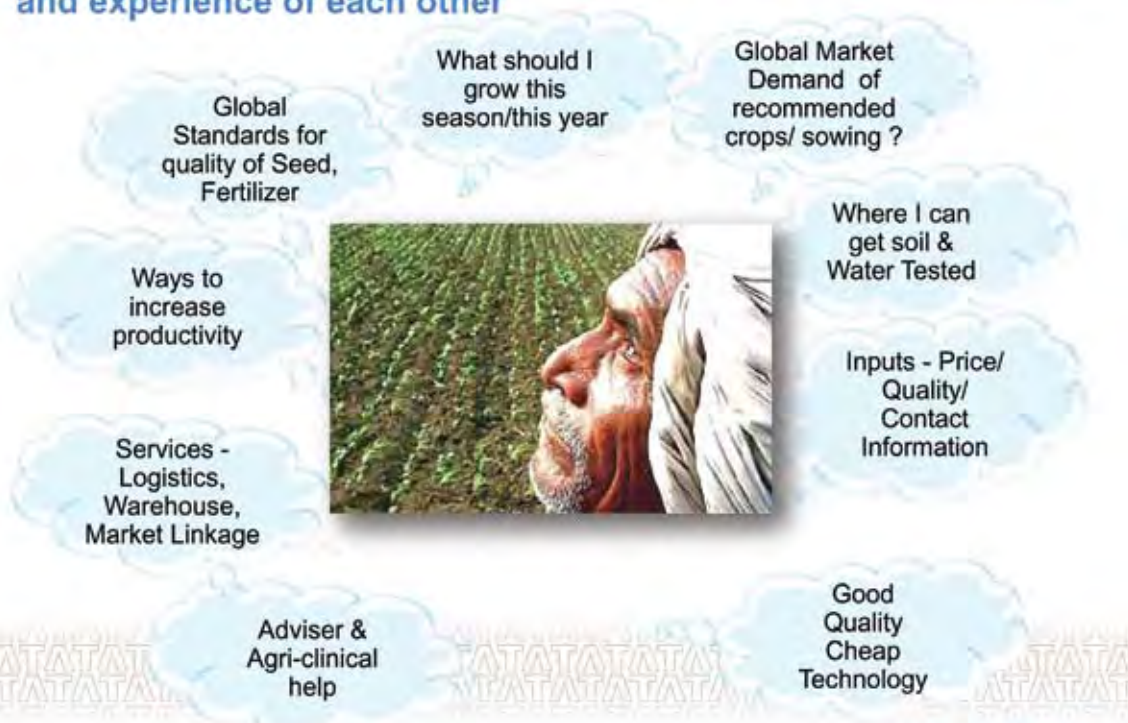


## How can Technology help the unreached and underserved?

TATA CONSULTANCY SERVICES

TCS Confidential

### Farmers face various issues and have to depend on knowledge and experience of each other



TATA CONSULTANCY SERVICES

14

TCS Confidential



## mKRISHI – Mobile Based Agro Advisory System

mKRISHI...

*Attempts to reduce the last mile gap between farmers and their eco system partners such as Agriculture experts, Markets, Government officials, Banks etc.*



TATA CONSULTANCY SERVICES

15

TCS Confidential

## mKrishi



- From a mere voice device
- To a multi-function system



- Micro region weather information and advice.
- Advice on fertilizer and pesticide usage
- Market information

TATA CONSULTANCY SERVICES

- 16 -





Thank You

[tanmoy.chakrabarty@tcs.com](mailto:tanmoy.chakrabarty@tcs.com)



**TATA** CONSULTANCY SERVICES

TCS Confidential



## Web for Everyone

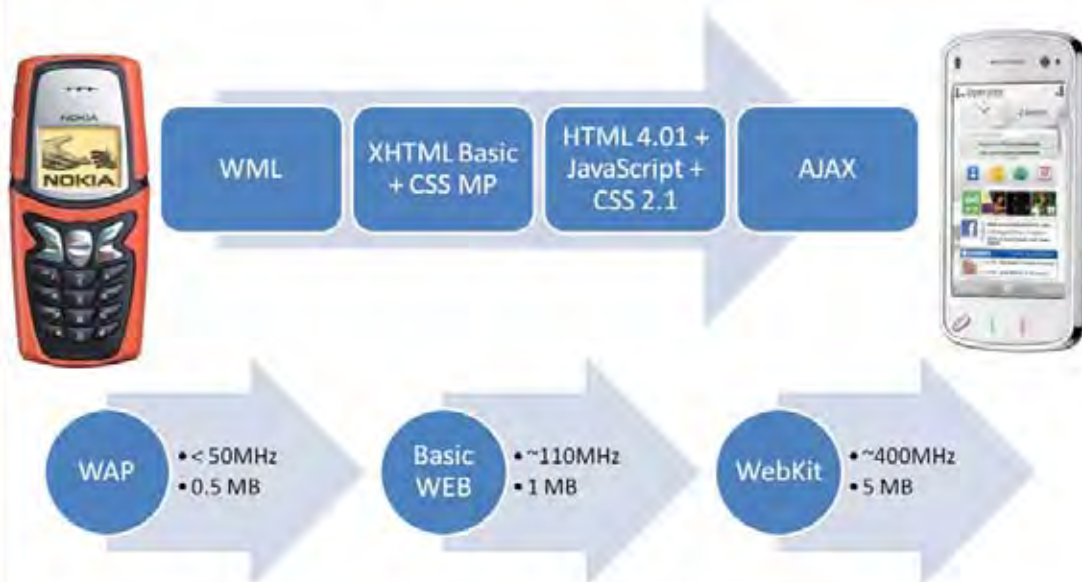
### Data enabling Mass Market Phones

C. Kumar

kumar.c@comviva.com

May 2010

## Mobile Browser Evolution



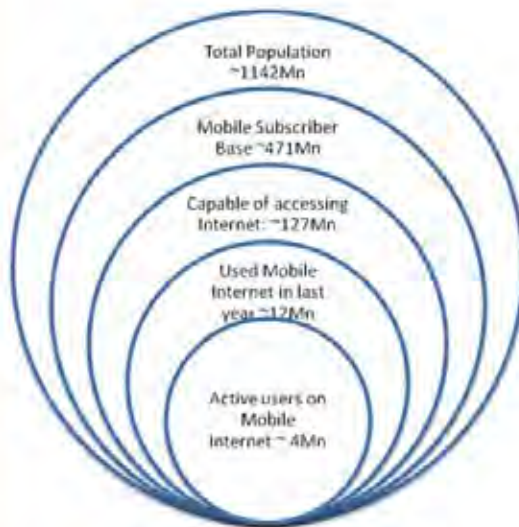
WebKit is becoming the de-facto standard for the Mobile Browser

2



## Mobile Internet Growth in India

comviva



Source: TRAI Sep. 2009 & I-Cube Estimates

### CHALLENGES

- Handset Performance – clock speed, memory
- Usability – smaller displays; does not have good Mobile Browser that adapts the Internet Content
- Language support – both on the device and content on the Internet
- Difficulties of launching browser and entering URLs
- Higher power consumption – requires more frequent charging of batteries
- Data Pricing – flat rate is getting adopted by most of the Mobile Operators

3

## How to overcome?

comviva

### Mobile Phone

- Solution should be adapted to the existing Mobile Phone without any upgrade for the CPU or additional memory.

### End User

- Usability & relevant content should be delivered.
- User needs to key-in less information – solution should come with pre-configured information.
- Power Consumption should be low.

### Developer

- Need to use the existing tools & technology – no learning required.
- Should use the content available on the Internet.

4



What is Mobile Widget?  
Does Mobile Widget address these issues?

## Mobile Widget

- Mobile Widgets are single purpose applications built using WEB technologies using the Mobile Internet.
- Mobile Widget Runtime (MWRT) is used to execute the Mobile Widgets on the phone.
- Success factor
  - NGPay
    - Similar to Mobile Widget with single purpose application to do the Mobile Commerce.
    - It allows to add more services within the single application.
    - Has around ~700000 users till date.



## Mobile Widget for mass market phones

comviva

Challenge	Description	Status
Content	Mobile widgets are single purpose application; can deliver right content with more graphics	😊
URL's	The user need not launch the browser and enter the URL's. Mobile Widgets are pre-configured.	😊
Usability & Performance	Performance will not be acceptable on the mass market phones as MWRT needs higher clock speed and increased memory	😞
Power Consumption	Mobile Widgets are powered by JavaScript and in order to compile and interpret at the device side it needs lots of power; but not as high as Mobile Browser.	😞
Mobile Price (BOM)	Good MWRT engine needs at least 4~5 MB of memory with processor speed (> 250 MHz). This shall increase the mobile phone price.	😞
Data Pricing	Data Pricing: Mobile Operator can offer better pricing as the data consumption over Mobile Widgets shall be very less compared to Mobile/Thin Browsers.	😊

7

comviva

How to take the benefits of Mobile Widgets to the mass market phone?

8



## Mobile Widget using Client-Server

comviva



- Mobile Widgets are deployed on the server and executed at the server side using the Mobile Widget Runtime.
- Server can act as a mobile application stores for the submission, certification and subscription.
- Thin client residing on the mobile device shall render the content.
- Client on the Mobile Device also interacts with the Mobile device features like Contact, Camera, Messaging and so on.

9

## Mobile Widget using Client-Server

comviva

Challenge	Description	Status
Content	Mobile widgets are single purpose application; can deliver right content with more graphics	😊
URL's	The user need not launch the browser and enter the URL's. Mobile Widgets are pre-configured.	😊
Usability & Performance	Thin Client software does not need high processor speed and it can be easily adapted to the mass market phones without any increase in the memory requirement.	😊
Power Consumption	JavaScript 's are executed at the server side and only the content are rendered on the client side. This shall reduce the overall power consumption.	😐
Mobile Price (BOM)	This solution can be easily adapted to the existing GPRS class phones without any additional memory/processor speed.	😊
Data Pricing	Data Pricing: Mobile Operator can define the flexible pricing policy based the Mobile Widget subscription.	😐

10



## W3C Role

comviva

- WEB Widgets was introduced by Apple Dashboard, popularized by Yahoo and mass adopted on Windows Vista.
- W3C widget shall help to standardize the interface for the Mobile Widget
- As part of the MWI – W3C can also provide certain guidelines similar to Mobile OK; so that it can be easily adapter to the low end mobile device as well.

11

## Conclusion

comviva

- An amazing opportunity to enable the first time Mobile Internet users to experience the Mobile Internet; and to help it evolve using Mobile Widgets.
- WEB for Everyone and for Every Mobile Device.



12



## About Comviva



- Comviva leading provider of end-to-end mobile VAS solutions in emerging markets globally.
- Comviva solution reaches 1 in 3 subscribers in emerging markets.
- Based in India with presence in Asia, EMEA and America.
- WebAxxn CS – Telco Grade Comviva MWRT Platform that supports Mobile Widget CS.
  - Tested on 250+ devices
  - It can be supported on any GPRS enabled device with MIDP2 support
  - It can be embedded on any low end GPRS device

13



[www.comviva.com](http://www.comviva.com)

Thank You





## Web for All – Indic languages perspective



Manish Bhargava, Google Inc. USA  
W3C Conference, New Delhi - May 6th, 2010

## Google's Mission





Country: United States

Language: English (US)



**Google WebSearch**  
World's most used web search



YouTube



Book Search



Gmail & Talk



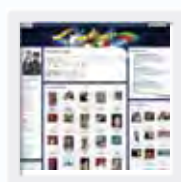
Earth



**AdSense for Search**  
Monetize search results pages



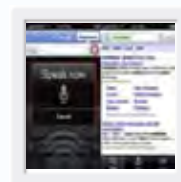
Maps



Orkut



Local



Voice Search



**AdSense for Content**  
Monetize content pages



Blogger



News



Picasa



Mobile

*Organizing all of the worlds information and making it universally accessible and useful*



## I18N/L10N – “Think Globally, Act Locally”

Google search in 150+ TLDs (Afghanistan - Zimbabwe)





# Addressing challenges to widespread adoption of Indic web in India

## International Comparison – Top 10 Spoken languages & Internet languages

Top 10 Languages Spoken in the World		Other Top Indian Languages	
Language	No. of Speakers	Language	No. of Speakers
Chinese	1.05 Billion	Telugu	75 Million
English	510 Million	Marathi	71 Million
Hindi	490 Million	Tamil	77 Million
Spanish	420 Million	Gujarati	46 Million
Russian	255 Million	Malayalam	37 Million
German	229 Million	Kannada	44 Million
Arabic	230 Million	Odia	32 Million
Bengali	215 Million		
Portuguese	213 Million		
Japanese	127 Million		

(Source: <https://www.vistawide.com/languages/>)

2 Indian languages in top 10 spoken languages

No Indian languages in top 10 Internet languages

Top 10 Languages in the Internet		
Language	% of Internet Users	No. of Users
English	30.5%	430.8
Chinese	20.4%	276.2
Spanish	6.8%	124.7
Japanese	1.9%	94
French	6.1%	68.2
German	1.4%	61.2
Arabic	5.4%	59.8
Portuguese	3.6%	58.1
Korean	1.1%	34.8
Italian	0.9%	34.7
Others	21.8%	220.9

(Source: <https://www.vistawide.com/languages/>)



## Heterogeneous user base

Segment	Addressable Market	Characteristics	Needs
<b>Non-Resident Indian (NRI)</b> (outside India)	• 20M users	<ul style="list-style-type: none"> <li>• Mobile and PC connected</li> <li>• High PC / broadband penetration</li> <li>• Economically well off</li> <li>• Early adopters for offerings that Developed segment will later utilize</li> </ul>	<ul style="list-style-type: none"> <li>• Entertainment Services that offer popular Indian content (music, movies, culture, tourism, etc.)</li> </ul>
<b>Developed</b> (in India)	• 300M users	<ul style="list-style-type: none"> <li>• Mobile and PC connected</li> <li>• Mostly urban, representing young middle class</li> <li>• Language heterogeneity</li> </ul>	<ul style="list-style-type: none"> <li>• Lifestyle Services that drive convenience and productivity</li> </ul>
<b>Emerging</b> (in India)	• 600M users	<ul style="list-style-type: none"> <li>• Limited connectivity (e.g. kiosk based)</li> <li>• Mostly rural</li> <li>• 100k villages will have kiosks by 2012</li> <li>• Means of access and income heterogeneity</li> </ul>	<ul style="list-style-type: none"> <li>• Livelihood Services that enable income generation and access to vital services (e.g. government forms, health care)</li> </ul>

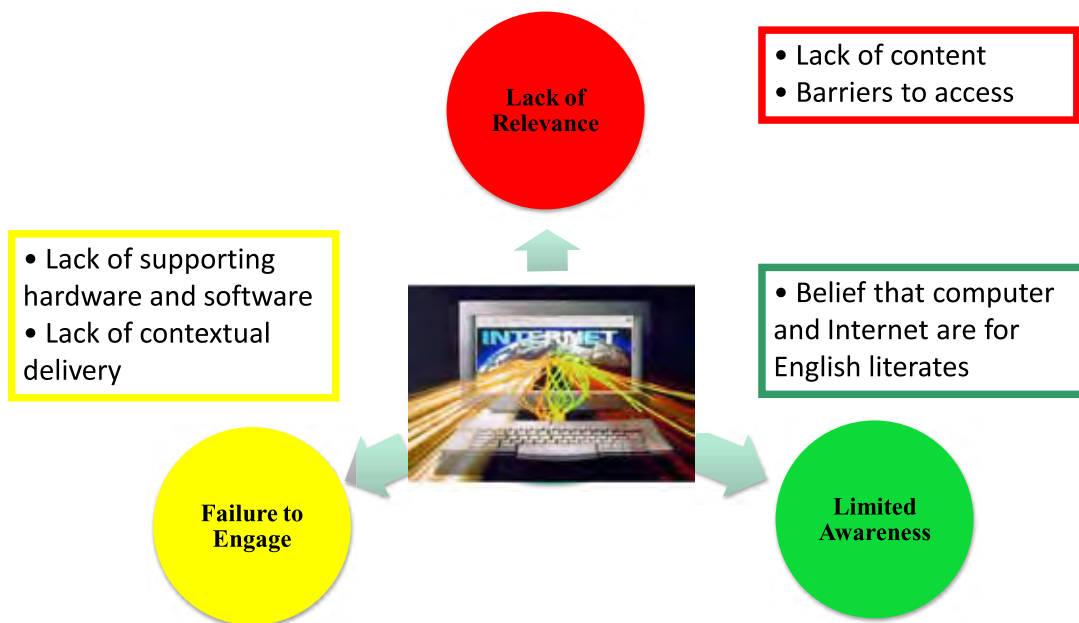
## Emerging Internet Users

- 5B People worldwide **not** on the Net
  - By current projections – only 50% of the world on Internet by 2030
  - India is a Microcosm
    - *Demographics*: 1.1 B people, \$1T GDP
    - *Penetration*: PC 2%; Internet 4%
    - *Internet Users*: 50M (40% yoy growth)
    - *Languages*: 22 official; 7% English proficient
- Internet is not as **Relevant**
  - *Absence of Content*: Local and Indic
  - *Barriers to Access*: Language, Mobile, Usability
  - *Lack of Contextual Applications*: 3<sup>rd</sup> Party Apps development
- Users not **Aware** of Internet's benefits
  - *Limited Awareness*: Overcome the current mindset that “computers” are for urban, english elite





## Challenges for Widespread Adoption of Indic web in India

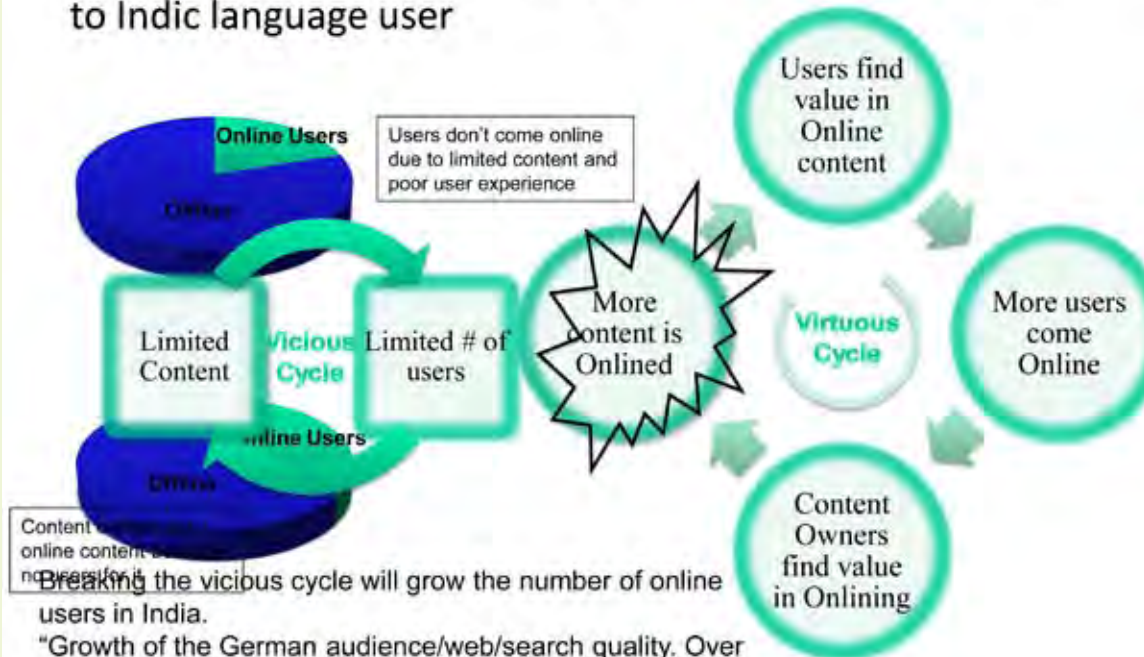


## Addressing Lack of Relevance



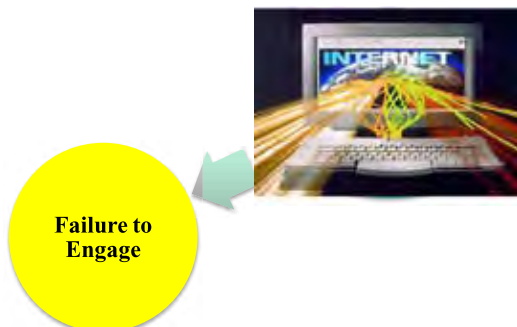


On-ling Content is Key to improve **Relevance** of the web to Indic language user



"Growth of the German audience/web/search quality. Over six months in 2000, German webpages went up from 25M to 50M. The corresponding growth in search quality and audience was rapid."

## Addressing failure to engage





## Language Tools to **engage** users

- Translation
  - Machine Translation
  - Text-to-Speech
  - Translation Element
- Transliteration
- Input Technologies
  - Virtual Keyboard
  - Indic IME
- Usability
  - Click-To-Search Experiment
  - Unified Language Settings
- Localized Apps



### Enable Translation of Internet Content

Content

Country:  Language:

Translation

The screenshot displays the Google Translate web interface. At the top, there are dropdown menus for 'Country' (set to India) and 'Language' (set to Hindi). Below these is a large 'Translation' button. The main content area shows the Google Translate logo and a search bar. On the right, there is a 'Translator Toolkit' button. The bottom section features a 'Machine Translation' label and a 'Hindi Dictionary' button. The dictionary shows a list of words in Hindi, including 'विद्या की पुरस्कार', 'पदार्थ की विज्ञान', 'पाठ्यक्रम की विद्यार्थियों की पदार्थ की पुरस्कार', and 'मुख्यमंत्री'.



# Create Content using Transliteration

Content

Country: India Language: हिन्दी

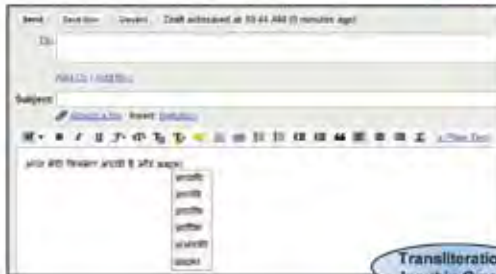
Transliteration



Orkut, Blogger Transliteration



Transliteration



Transliteration Input in Gmail



Script Conversion

# Make it easy to input vernacular language

Input

Country: India Language: हिन्दी

Input



Virtual keyboard in Hindi



Indic transliteration IME



## Make it easy for users

Guidance

Country: India Language: हिन्दी

Usability



## Localize Relevant Applications

Apps

Country: India Language: हिन्दी

Localized Apps





## Google Products to address these challenges

- Local Information
  - MapMaker, Google Places
  - OneBoxes
- Unstructured Content
  - Atlantis
- Translated Corpus
  - Localization Toolkit
  - Machine Translation
- User Generated Content
  - YouTube
  - Blogger with Transliteration



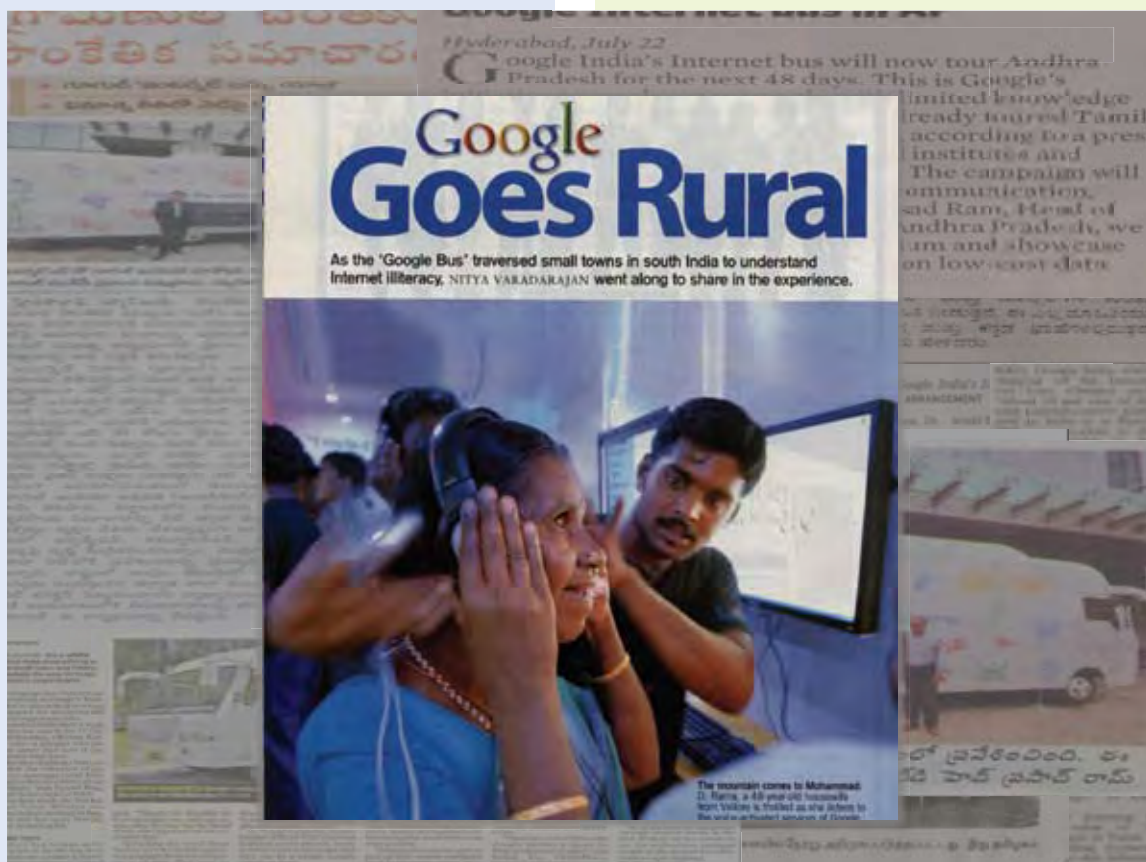
## Addressing lack of awareness





4	430,000
60	250
16,000	26,000,000

States, Cities, Subscribers, People,  
Articles, Reach





Awareness

# The Internet Bus Project

[www.google.co.in/internetbus/](http://www.google.co.in/internetbus/)



## Internet Bus Goals

Awareness

Empower people with information by raising awareness of the Internet





# Google's Internet Bus – Inside View

Awareness



Education	Entertainment	Communications	Information
<p>Explore links below பேற்றொன்று ஆறிய விதே கிளிக் செய்யவும்.</p> <p>Top Links</p> <p>Tamil Story Book தமிழ் கதையின் மணிகள்</p> <p>Tamil History தமிழ் வரலாறு</p> <p>Anna University</p> <p>Learn Computers</p> <p>NIIT</p> <p>Children's Education குழந்தைகளின் கல்வி</p> <p>Secondary Education மேல்நிலை கல்வி</p> <p>Higher Education உயர் கல்வி</p> <p>Jobs / Vocational Training தொழிலகல்வி/தொழில் பயிற்சி</p> <p>Google Internet Bus</p>	<p>Explore links below பேற்றொன்று ஆறிய விதே கிளிக் செய்யவும்.</p> <p>Top Links</p> <p>cricinfo</p> <p>Incredible India</p> <p>Cricket News</p> <p>Incredible India</p> <p>Indian Festivals</p> <p>Latest Releases</p> <p>Movies / Music திரைப்படங்கள்/இசை</p> <p>Culture கலாச்சாரம்</p> <p>Sports / Games விளையாட்டுகள்</p> <p>Travel பயணம்</p> <p>Google Internet Bus</p>	<p>Explore links below பேற்றொன்று ஆறிய விதே கிளிக் செய்யவும்.</p> <p>Top Links</p> <p>YouTube</p> <p>Share Videos</p> <p>orkut</p> <p>Orkut</p> <p>Government Services</p> <p>Google Google SMS Search</p> <p>Mobile மொபைல்</p> <p>Connect With Friends தங்கியுள்ள நோட் பிடித்துக்கொள்</p> <p>Email / Messaging எடுப்பித்தகல்பம் பரிமாற்றம்</p> <p>Contact Government அரசாங்கத்தோடு தொடர்பு கொள்ளுங்கள்</p> <p>Google Internet Bus</p>	<p>Explore links below பேற்றொன்று ஆறிய விதே கிளிக் செய்யவும்.</p> <p>Top Links</p> <p>தினகரம்</p> <p>Dinamalar தினமலர்</p> <p>Weather வானிலை</p> <p>Webdunia</p> <p>Webdunia செய்துமலியர் தமிழ்</p> <p>Google Google Tamil தமிழ் Google தேடல்</p> <p>News செய்திகள்</p> <p>Find Answers வினா... கண்டுபிடிப்புகள்</p> <p>Everyday Life ஆவற்றாட... வாழ்வுகள்</p> <p>Business வணிகம்</p> <p>Google Internet Bus</p>



## Addressing challenges for Widespread Adoption of Internet in India





## Automatic Speech Recognition – Research and Standards

**S. Umesh**

(with Raghavendra, Kishore Prahlad, Hema Murthy)

Department of Electrical Engineering  
Indian Institute of Technology Madras  
May 7<sup>th</sup>, 2010

S. Umesh (IIT-M)

W3C 2010

May 7<sup>th</sup>, 2010

1 / 21

### Outline

- Automatic Speech Recognition (ASR)
- ASR engines from academia
- ASR engines from industry
- Flexibility & Limitation of academia ASRs
- Existing Standards

S. Umesh (IIT-M)

W3C 2010

May 7<sup>th</sup>, 2010

2 / 21



## Automatic Speech Recognition (ASR)



- **ASR technology:** allows a computer to recognize words that a person speaks into a microphone or telephone. Convert the input speech into text.
- Articulators produce sounds which the ear conveys to the brain for processing.

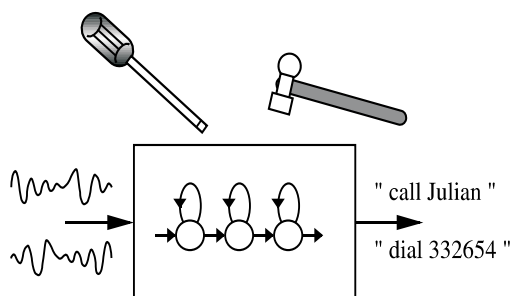
S. Umesh (IIT-M)

W3C 2010

May 7<sup>th</sup>, 2010 3 / 21

## Automatic Speech Recognition (ASR)

ASR - Convert Speech signal to words



- Most languages: only 50-60 distinct sound units make up the words
- Example :

and - sil /a/ /n/ /d/ sil  
 yes - sil /y/ /E/ /s/ sil  
 no - sil /n/ /ow/ sil

S. Umesh (IIT-M)

W3C 2010

May 7<sup>th</sup>, 2010 4 / 21



## Pronunciation Dictionary

Pronunciation Dictionary: Expand words to corresponding sounds

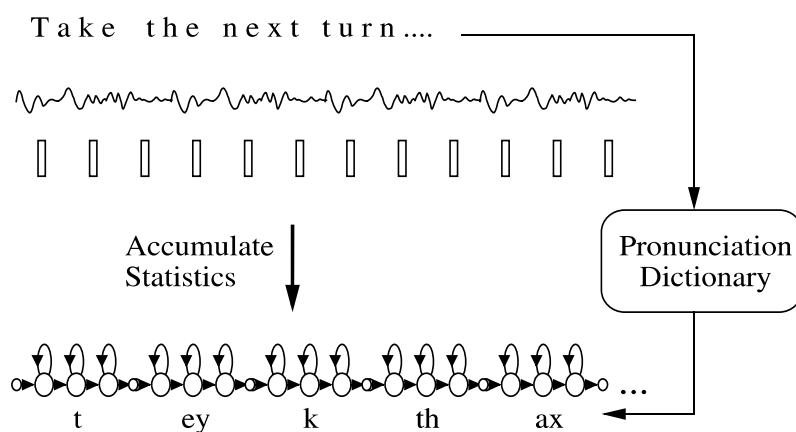
A	ah
A	ax
A	ey
CALL	k ao l
DIAL	d ay ax l
EIGHT	ey t
PHONE	f ow n
SEVEN	s eh v n
TO	t ax
TO	t uw
ZERO	z ia r ow

S. Umesh (IIT-M)

W3C 2010

May 7<sup>th</sup>, 2010 5 / 21

## Training of a Speech Recognition System



S. Umesh (IIT-M)

W3C 2010

May 7<sup>th</sup>, 2010 6 / 21



## Language Model (LM)

- Example:

"It's fun to recognise speech?"

"It's fun to wreck a nice beach?"

Although the "sound-sequence" may be similar, LM will tell us that the first sentence is more likely

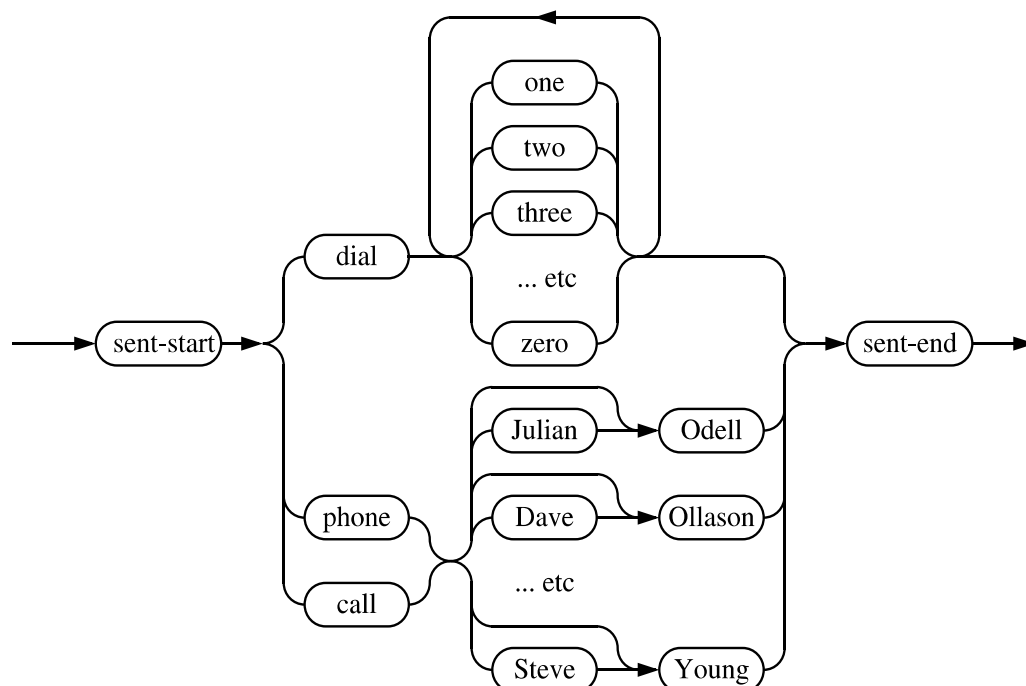
- Language models are used to restrict the combination of words
- Permissible words following each word are given explicitly in LM

S. Umesh (IIT-M)

W3C 2010

May 7<sup>th</sup>, 2010 7 / 21

## Grammar



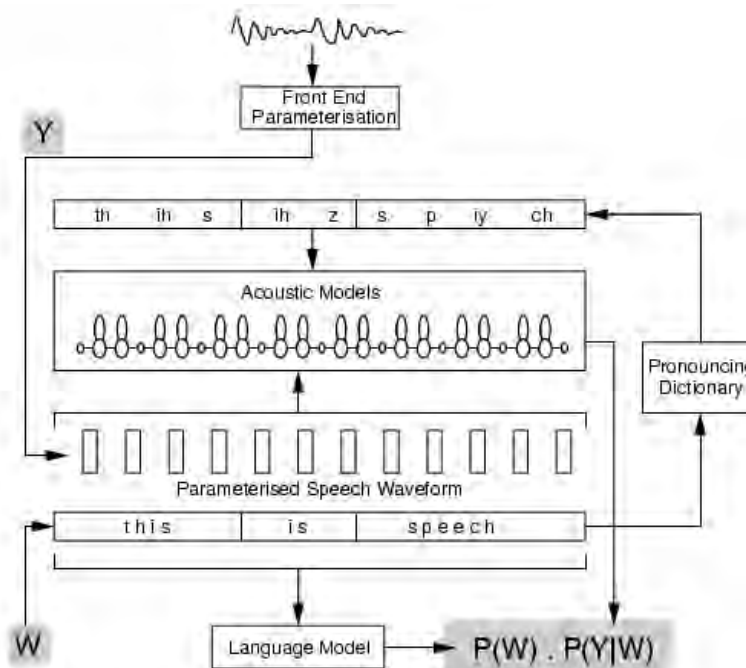
S. Umesh (IIT-M)

W3C 2010

May 7<sup>th</sup>, 2010 8 / 21



## Statistical View Point of ASR



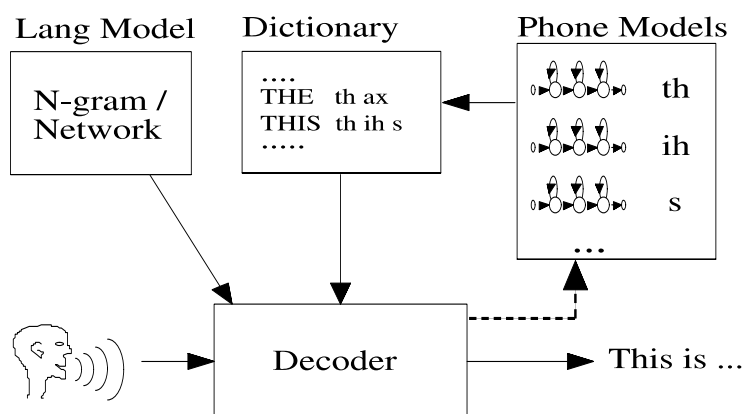
S. Umesh (IIT-M)

W3C 2010

May 7<sup>th</sup>, 2010 9 / 21

## Recognition

- Speech Recognition Grammar Specification (SRGS)
- Pronunciation Lexicon Specification (PLS)



S. Umesh (IIT-M)

W3C 2010

May 7<sup>th</sup>, 2010 10 / 21



## ASR Engines from Academia

- Sphinx (CMU)
- HTK (Cambridge University)
- SUMMIT (MIT)
- SONIC (University of Colorado)
- Julius (CSRC, Japan)
- CSLU (OGI school of Science and Engineering)

S. Umesh (IIT-M)

W3C 2010

May 7<sup>th</sup>, 2010

11 / 21

## ASR engines from Industry

- Loquendo ASR
- Dragon Naturally Speaking
- TeliSpeech Recognizer
- IBM ViaVoice
- MacSpeech
- Simmortal Voice
- e-Speaking
- VoiceFinger
- LumenVox Speech Engine

S. Umesh (IIT-M)

W3C 2010

May 7<sup>th</sup>, 2010

12 / 21



## Flexibility & Limitations of Academia ASRs

- Very Flexible: Lot of freedom to make changes in different modules
- Geared towards promoting research in different modules
- Interoperability between ASRs is difficult
  - An module working Sphinx cannot be easily made to work in HTK
  - Input, output specifications differ between ASRs
  - Storage formats are very different
- To use ASR in various applications, system should support standards.
  - Allow easy interoperability
  - To have a plug-n-play ASR
- Industry engines follow standards but still do not allow easy inter-operability between various industry engines.

S. Umesh (IIT-M)

W3C 2010

May 7<sup>th</sup>, 2010

13 / 21

## Conclusion

- Standards exist for many modules of ASRs but not all
- For some applications, current standards may not provide enough flexibility
- Academic ASRs do not follow standards but provide flexibility
- Implementation aspects of Standards in ASR differ significantly from research aspects of ASR

S. Umesh (IIT-M)

W3C 2010

May 7<sup>th</sup>, 2010

14 / 21



## Standards

- The various kinds of standards over internet, distributed system and desktops are as follows.
  - Internet; Voice Extensible Markup Language (VXML)
  - Distributed environment; Media Resource Control Protocol (MRCP)
  - Desktop; Speech Application Programming Interface (SAPI)
  - ASRs developed by the academia does not follow above standards where as industry developed systems (such as supports most one or two
- Microsofts SAPI
- W3C Standards
  - Voice XML
  - Media Resource Control Protocol (MRCP)
  - Speech Recognition Grammar Specification (SRGS)
  - Pronunciation Lexicon Specification (PLS)

S. Umesh (IIT-M)

W3C 2010

May 7<sup>th</sup>, 2010

15 / 21

## VXML

- VoiceXML (VXML) is the W3C's standard XML format
- Specify interactive voice dialogues between a human and a computer.
- A kind of programming language that helps computers and other devices operate through telephone lines.
- Allows voice applications to be developed and deployed in an analogous way to HTML for visual applications.
- As HTML documents are interpreted by a visual web browser, VoiceXML documents are interpreted by a voice browser or IVR.
- VoiceXML has tags that instruct the voice browser to provide speech synthesis, automatic speech recognition, dialog management, and audio playback.

S. Umesh (IIT-M)

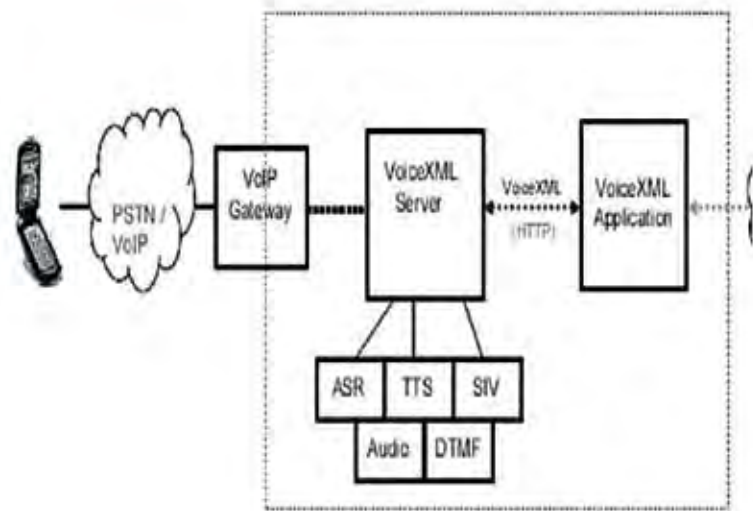
W3C 2010

May 7<sup>th</sup>, 2010

16 / 21



## Voice XML Application Architecture



S. Umesh (IIT-M)

W3C 2010

May 7<sup>th</sup>, 2010

17 / 21

## Voice XML and MRCP

- Media Resource Control Protocol (MRCP) helps to talk to different speech engines (TTS, ASR, Speaker id) in an efficient fashion.

S. Umesh (IIT-M)

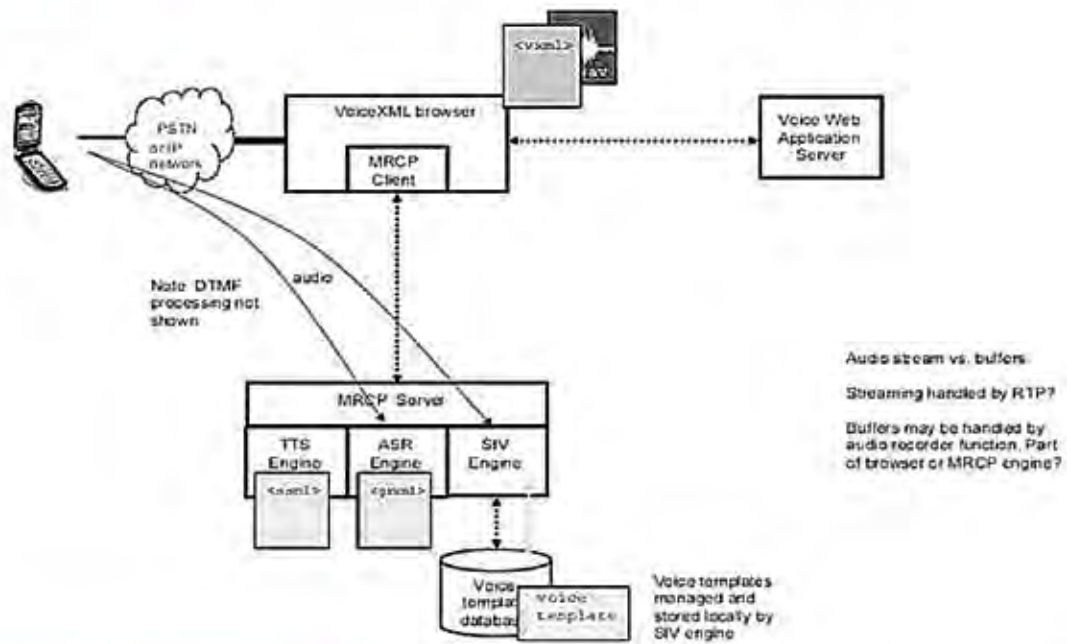
W3C 2010

May 7<sup>th</sup>, 2010

18 / 21



## VXML, MRCP



S. Umesh (IIT-M)

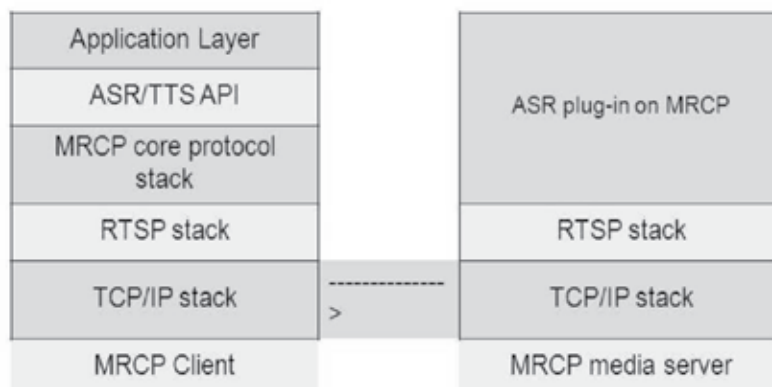
W3C 2010

May 7<sup>th</sup>, 2010

18 / 21

## MRCP

- MRCP is designed to provide a mechanism for a client device requiring audio/video stream processing to control processing resources on the network.
- The MRCP protocol defines the requests, responses, and events needed to control the media processing resources.
- Architecture



S. Umesh (IIT-M)

W3C 2010

May 7<sup>th</sup>, 2010

20 / 21



## SAPI

- The SAPI is an API developed by Microsoft to allow the use of speech recognition within Windows applications.
- In general all versions of the API have been designed such that a software developer can write an application to perform speech recognition and synthesis by using a standard set of interfaces, accessible from a variety of programming languages.
- In addition, it is possible for a 3rd-party company to produce their own Speech Recognition engine or adapt existing engines to work with SAPI. In principle, as long as these engines conform to the defined interfaces they can be used instead of the Microsoft-supplied engines.



## A Framework for Secure Communication using Hindi for Web-based and Mobile Applications

### Authors :

**Dr. Saibal K. Pal**  
*Scientific Analysis Group,  
DRDO, Delhi India*

**Sarvesh Kumar**  
**Sarvejeet Kumar**  
**Mukesh Kumar**

*Department of computer Science  
University of Delhi, Delhi India*

## Need for Translation and Security of Regional Language

- **Translation**

As we know India is a multilingual country, so communication takes place in many regional languages .

The translation of a regional language into corresponding English script will provide a way to connect people through their native languages.

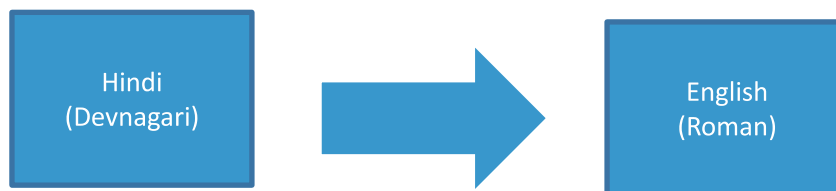
- **Security**

Many a times it is required to secure a communication in regional language over the networks.

For example to secure an e-mail in a regional language.



## Script Translation



How ??

## Rule-based Script Translation

- DEVNAGARI
  - Phonetics based script
  - Words are largely written according to phonetics(pronunciation)
  - Possibility that a written word is pronounced differently is very less
  - Does not require memorizing spellings

*These characteristics of Devnagari Script makes it EASY to convert it into Roman*



## Rule-based Script Translation (contd.)

- The pronunciation of Hindi(Devnagari) words is not continuous, it takes momentary stops in between.

e. g. सरगम (SARGAM)

*This Characteristic of Devnagari Script makes it DIFFICULT to convert Devnagari into Roman.*

## Rules used in Translation

- Anusvaar rule**

The occurrence of Anusvaar(ँ) in a Hindi word produces two sounds 'ँ' and 'ँ' depending upon the next consonant in the word.

If the next consonant is from 'प वर्ग' then Anusvaar produces sound of 'ँ' otherwise sound of 'ँ'.

e.g. संबंध = SAMBANDH

- Laghu swar and Deergh swar rule**

e.g.

कमला = Kamla OR Kamala (inappropriate but correct according to std. Hindi)

In this example there is a DEERGH SWAR on 'ल' hence 'a' is not placed between 'm' and 'l'.



## Rules used in Translation (contd.)

- **Rearrangement of vowel marks ( MATRA)**

Rearranging REF ( <sup>ˆ</sup> )

e.g.      सर्वजन      is written as      स व <sup>ˆ</sup> ज न  
             सर्वजन      is pronounced as      स <sup>ˆ</sup> व ज न (SARVJAN)

Rearranging ( ि )

e.g.      किताब      is written as      ि क त ा ब  
             किताब      is pronounced as      क ि त ा ब (KITAB)

Here the Matra( ि ) is placed at the position where it's pronunciation occurs.

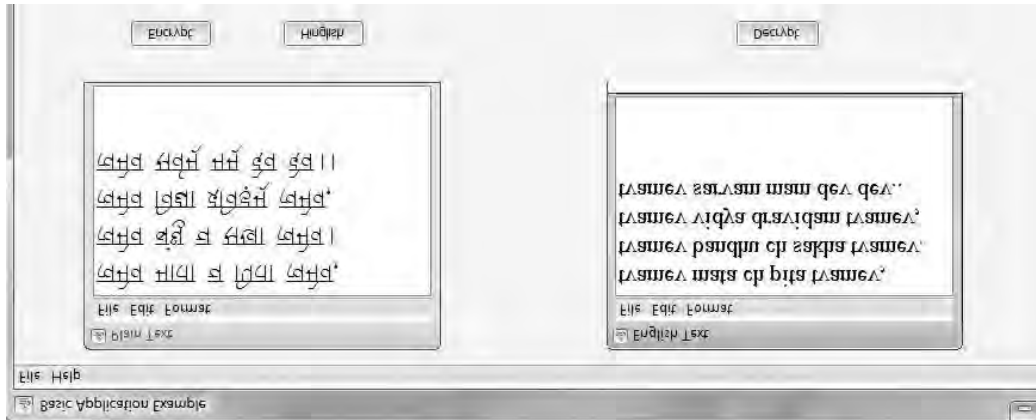
## Screen-shot



**Hindi (Devnagari) to English (Roman)**



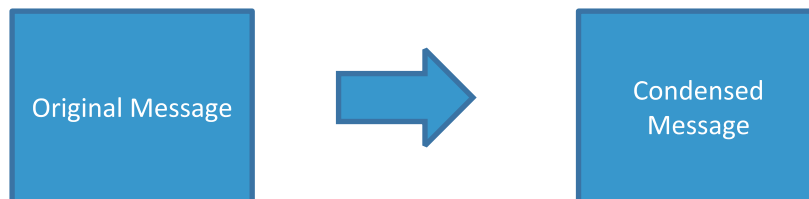
## Screen-shot



Sanskrit (Devnagari) to English (Roman)

## Rule-based Condensation of Message

- We can condense the message according to rules of texting languages as we do in **SMSs**.





## Rule-based Condensation of Message (contd.)

- Abbreviations and Symbols can be used for certain frequent words in the message.

e.g. **mob.** for **mobile**  
**@** for **at**  
**&** for **and**  
**2day** for **today**

etc.

- Vowels can be omitted in a word in such a way that reduced word convey same meaning as that of original. Repetition of consonants can also be omitted.

e.g.

**shortest** → **shrtst**  
**good** → **gd**  
**something** → **smthng**  
**message** → **mssg** → **msg**

## Screen-shot



English to SMS







# Thank You

**Contact:**

Mobile : +919650944960

Email : [stonesarvesh@gamil.com](mailto:stonesarvesh@gamil.com)





## Indic Text Display Issues on Digital Devices

Delivering Unmatched User Experience



breaking language barrier



All images, brand names, logos, trademarks used in this document are copyrights of respective owners. Reverie does not claim any rights over any 3<sup>rd</sup> party property. The images used are for demonstration purposes only.


copyright 2009 | confidential [www.reverie.co.in](http://www.reverie.co.in)

### The Choice

Our bias for English



English



Indic

copyright 2009 | confidential 



## India's Linguistic Preference

### Facts & Myths

- 10% English Literacy
- Top 10 Newspapers - Regional
- Each of Top 4 Regional Newspaper = 3x Top English Newspaper
- Star TV Avatar 1 - 100% English (Unsuccessful)
- Star TV Avatar 2 - 100% Hindi (Successful)

**India is a Multi-Lingual Country**

copyright 2009 | confidential



## The Force

### In digital media



**Print Media**



**Digital Media**

copyright 2009 | confidential





## The Display Issues & Problems

The beginning of user experience

### Challenge 1

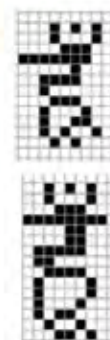
### Unacceptable Quality of Rendering & Fonts

copyright 2009 | confidential

REVERE  
TECHNOLOGIES

## Current Quality of Display on Mobiles

Digital media yet to mature



copyright 2009 | confidential

REVERE  
TECHNOLOGIES



## Current Quality of Display – Rendering Issues

### Set-Top-Boxes



copyright 2009 | confidential

REVERE  
TECHNOLOGIES

## Current Quality of Display – Composition Issues

### Set-Top-Boxes



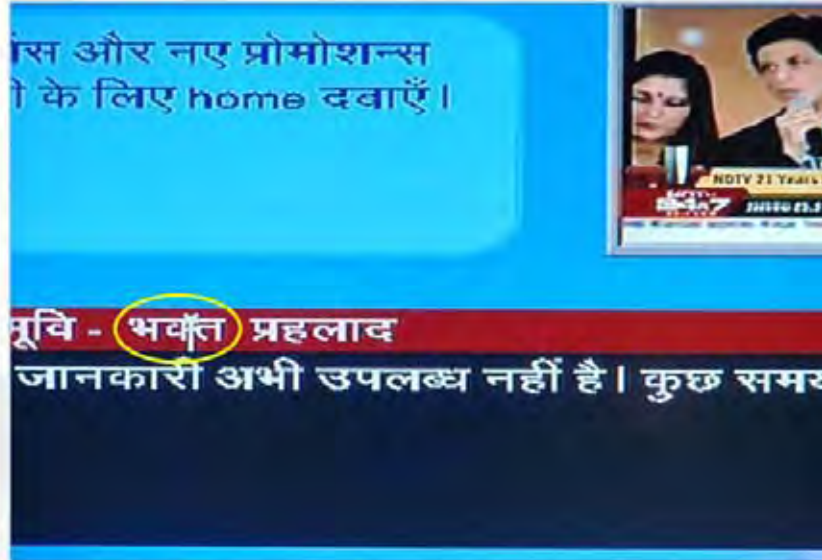
copyright 2009 | confidential

REVERE  
TECHNOLOGIES



## Current Quality of Display – Fonts Issues

### Set-Top-Boxes



copyright 2009 | confidential

REVERE  
TECHNOLOGIES

## The Display Challenge

### The "INDLISH" Way

A+P+A+R+T+M+E+N+T = APARTMENT (not APATRMENT)

But

अ+प+अ +रु +ट +म+ए +न् +अपाट्मेंट

S+P+O+R+T+S = SPORTS (not SPOTRS)

But

स+प+ओ +र+ट् +स = स्पोर्ट्स

Indic Scripts need 100% accurate rendering

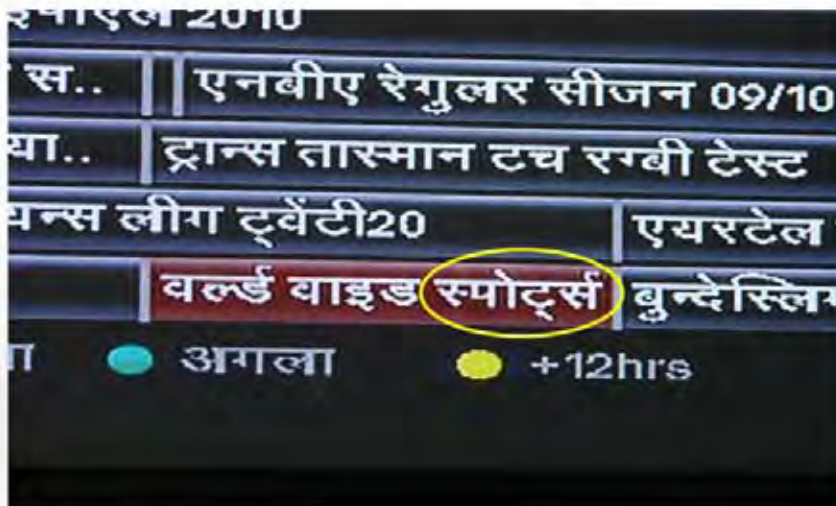
copyright 2009 | confidential

REVERE  
TECHNOLOGIES



## Current Quality of Display

Set-Top-Boxes



copyright 2009 | confidential



## Our Approach to The Display Challenge

Usability is Acceptability

- Guarantee 100% accurate rendering
- Ensure legibility as good as or better than in print media
- Accommodate all scripts in entry level devices

Get the BASICS right

copyright 2009 | confidential



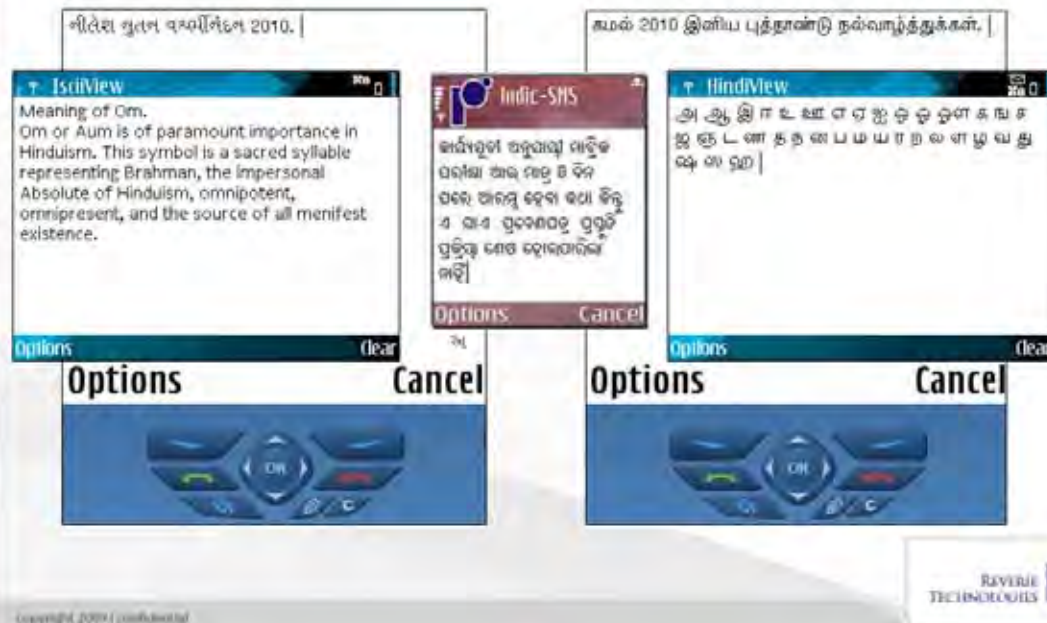






## Our Display Framework- Font Aesthetics

### Other Languages – Mobile Phones



## Our Display Framework

### Set-Top-Boxes

धूम मचा ले .. (धूम 2) 2006  
 आज्ञा नचले (आज्ञा नचले) 2007  
 आँखों में तेरी अजब-सी ...(ओम शांति ओम) 2007  
 बरसो रे, मेघा-मेघा ... (गुरु) 2007  
 चकदे चकदे इंडिया (चकदे इंडिया) 2007  
 मौजा ही मौजा ... (जब वी मेट) 2007  
 सजना जी, वारी-वारी जाऊँ जी मैं ..(हनीमून ट्रैवल्स)



## Our Display Framework

Set-Top-Boxes

ரிமசிம-ரிமசிம, ருமசும-ருமசும (1942 ஏ லவ  
தூம மசா லே .. (தூம 2) 2006  
ஆஜா நசலெ (ஆஜா நசலெ) 2007  
ஆங்கோங் மேங் தேரீ அஜப-ஸ் ...(ஓம் ஷாங்  
பரஸோ ரே, மேகா-மேகா ... (குரு) 2007

copyright 2009 | confidential



## The Display Issues & Problems

The beginning of user experience

### Challenge 2

### Unicode Implementation for Languages

copyright 2009 | confidential









## The Unicode Challenge – Proposed Approach

### Language vs. Script

- Unicode script code pages are almost exhaustive for a **“script”**
- Support for a language is significantly simpler
- 3GPP Indic Language Tables also adopt the simplicity

**“LANGUAGE”** is different from a **“SCRIPT”**



## What Do We Do Differently?

### English is **“NOT”** an Indic Script

Properties	English	Indic
No. of Characters	26	>50
Nature of script	Linear	Non-linear
Script Code	ASCII	Unicode
Conjuncts	Non-existent	Backbone
Case Handling	Lower / Upper	Non-existent
Language	Non Phonetic	Phonetic

**Indic + English ≠ “INDLISH”**





## Summary

### Key Messages

- Give Indian Language its due in digital mediums
- Adopt “**Indic Centric**” and not “**Indlish Centric**” approach
- Address all aspects of solution to ensure “**COMPLETENESS**”
  - Rendering & Fonts
  - Text Entry / Edit
  - Unicode Implementation
  - Standardization
- English IS an Indian language. Adopt its merits for Indic Scripts
  - Simple
  - Standardized
  - User Friendly

**USABILITY and not just AVAILABILITY**



## About Us

### Pioneers in Indian Language Computing

- **25+ years experience in complex Indic and South Asian Scripts**
  - Devanagari, Bengali, Assamese, Gujarati, Punjabi, Oriya, Telugu, Tamil, Kannada, Malayalam, Sinhalese, Tibetan, Bhutanese, Urdu, Perso-Arabic
- **Founder pioneer of digital type design for Indic Scripts**
  - Headed font design & standardization at C-DAC for 10 years
- **India's 1st & only TV platform with 11 Indic scripts**
  - As early as in 2001
- **India's 1st Indic Scripts Solution for HDTV**
  - In 2007
- **Designed more than 200 digital Indic Scripts typefaces (Since 1988)**
  - LEAP, ISM
  - Movie Sub-titling in all Indian languages - LIPS, MOVE suite of products
- **Representation in Indian Ministry for defining Indic standards**
  - Since 1988
- **Leading Indic standardization efforts for mobile phones**
  - Through CEVIT in collaboration with COAI, TRAI and Industry Leaders



Copyright 2009 | Confidential



## Industry Recognition

### Prestigious Awards



FIE Foundation National Award '93



CDAC Award In '95

Copyright 2009 | Confidential



Thank You

Copyright 2009 | Confidential







# Internationalization & Localization : Indian Perspective and requirements

By :  
Swaran Lata  
Country Manager, W3C India Office  
6, CGO complex, Electronics Niketan  
New Delhi  
E-mail : slata@mit.gov.in



Internationalization and localization are subsets of globalization

Taking a product and making it linguistically and culturally appropriate to the target locale (country/ region and language) where it will be used and sold"



Localization

Process of generalizing a product so that it can handle multiple languages and cultural conventions without the need for re-design.



Internationalization







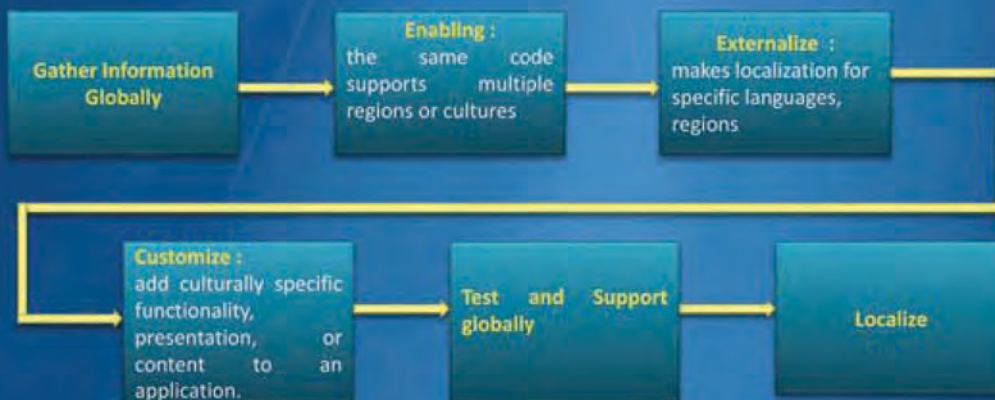
## Localization Vs Internationalized



- Designing and developing in a way that removes barriers to localization or international deployment.
- Providing support for features that may not be used until localization occurs.
- Enabling code to support local, regional, language, or culturally related preferences.
- Separating localizable elements from source code or content, such that localized alternatives can be loaded or selected based on the user's international preferences as needed.
- It can be localized quickly.



## The Internationalization Approach to Globalization







## Complexity of Indian Scenario

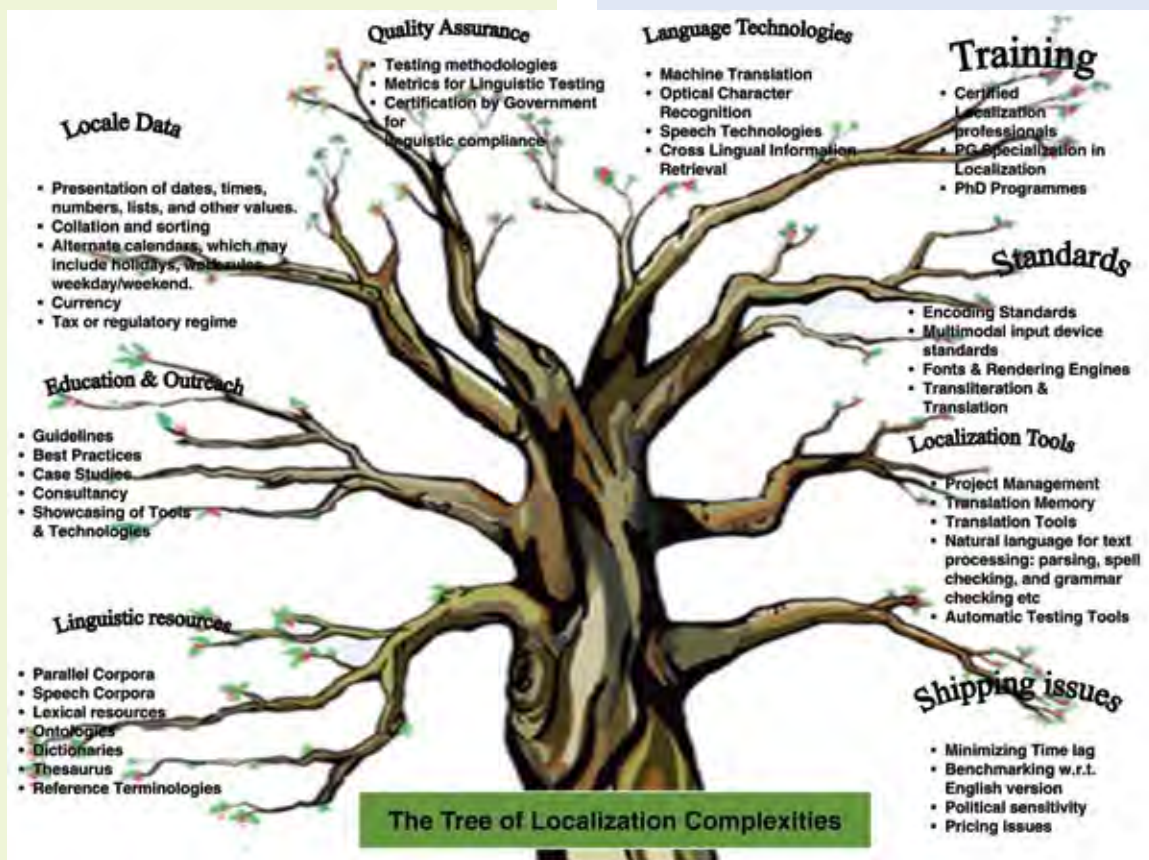


- India is Multilingual Multi script Country with 22 languages and 11 scripts; population over 1 Billion
- Less than 5 percent of people can read & write English. Over 95 percent population is deprived of the benefits of English-based Information Technology

### Issues regarding Indian Languages

- Orthography – Spelling issues
- Pronunciation – may be directly mapped but not always
- One script-many languages
- Many languages – one Script





## Internationalization Activity in W3c



- The W3C Internationalization (I18n) Activity works with W3C working groups to make it possible to use Web technologies with different languages, scripts, and cultures.
- It is to ensure that W3C's formats and protocols are usable worldwide in all languages and in all writing systems.
- The Internationalization (I18n) Activity statement explains concepts relating to internationalization, as well as the current situation and the role within the W3C of the Internationalization Activity.





## W3C Internationalization Resources



- **Internationalization of Web design & Applications**
  - Character Model for World Wide Web
  - Authoring Techniques for XHTML & HTML
  - Authoring CSS
  - Unicode in XML
- **Internationalization of Web Architecture**
  - Language tags and Local Identifiers
  - Internationalization Tag Set
- **Internationalization of XML**
  - Best practices for XML Internationalization
- **Internationalization of Web Services**
  - Language tags and Local Identifiers for World Wide Web



## Activities to be undertaken by W3C India Office



- ❖ **Internationalization**
  - ❖ Internationalization Tag Set
- ❖ **Web Design and Applications**
  - ❖ Styling
  - ❖ Html
  - ❖ Xhtml
  - ❖ Wai
- ❖ **Web Architecture**
  - ❖ XML
- ❖ **Semantic Web**
  - ❖ OWL and RDF
- ❖ **XML Technology**
  - ❖ XML associated standards
- ❖ **Web of Services**
  - ❖ SOA
- ❖ **Web of Devices**
  - ❖ Mobile Web Initiative
  - ❖ Voice
- ❖ **E-Government**
  - ❖ Use cases







## MOBILE WEB



### Challenges :

- Adopting right encoding scheme
- Availability on handsets
- Usability of Mobile Web Browser
- Web support of all Indian languages
- Study on specific requirements for Indian languages for W3C Mobile Web Best Practices
- Must support standards and specifications
- Access to all handset features



## MOBILE WEB



### Issues on Mobile Web

- Character Encoding
- Bandwidth and Cost
- Presentation Issues
- Input
- Device Limitations
- Lack of standardization
- Fonts
- Backward Compatibility with Legacy Devices
- Lack of standardization
- Rendering Issues

### Messaging Issues

- Lack of availability for all characters.
- There is no guarantee that a message encoded will be displayed properly at the receiving terminal.
- Issue of Multiple Script -one language not addressed.
- Standardization of glyph support, syllable composition logic is also an important aspect and is dependent on the implementation level of handset manufacturer.
- Legacy Systems





## MOBILE WEB



### Issues in Mobile Keypads

- Multi-tap issues
  - Too many taps per Key
- Dictionary Based issues
  - Difficult to learn and operate for the target segments.
  - Different spelling for मुर्ती, मुर्ती, मूर्ति even मुरथी, many permutations. Which is the one to be mapped

#### Hindi Alphabets mapping in Mobile

2	अ आ इ ई उ ऊ ऋ ॠ
3	ए ऐ ओ औ अः
4	क ख ग घ ङ
5	च छ ज झ ञ
6	ट ठ ड ढ ण
7	त थ द ध न
8	प फ ब भ म
9	य र ल ळ व श ष स ह

to know which char is on which



## MOBILE WEB



### Suggestions

- In terms of internationalization, operators must support appropriate character encoding on the signaling channel which would allow all characters of the world to be represented.
- Need of investigation and study of major issues for enabling Mobile web in Indian languages
- Standardization of mobile media also required to be addressed taking into consideration of specific requirements of each of Indic languages.

### Road Map :

- Character encoding as per Unicode Standard
- Enable Mobile Web in Indian languages
- Initiation of study for Mobile Web Best Practices 1.0 with respect to requirements for 4 Indian languages : Hindi, Bangla , Marathi, Tamil









## Cascading Style Sheet



- Underlining of characters
- There is some examples of Indian languages in which Matra's are not readable due to underlining of characters

**Hindi -** अन्य भाषाओं में भी अनुवाद

**Punjabi-** ਬਾਬੂ

**Bengali-** তাই পুরোনো আর্কাইভ একটু ওলট পালট।

**Guajarati -** સરદાર ગ્રજરી

**Marathi-** मराठी मुला मुलींची नावे

**Tamil-** நீளற்குமிழி யிளமை நிறைசெல்வம்

**Telugu -** శ్రీలం ప్రజ్ఞా TV9 ప్రోగ్రాం " డ్యూమి ఇన్ డెంజర్ " పార్ట్



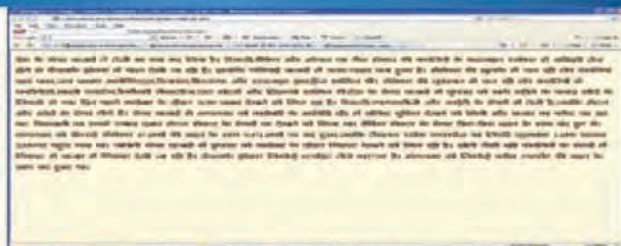
## Cascading Style Sheet



- Vertical arrangements

चौ	व	or	व	or	वक्	श	श	
द	का		क्		ता	कि	or	क्
			ता					ति

- Formatting issues :
  - Horizontal justification



- Bullets and Numbering

अ )	U+0905	अ )	U+0A78
आ )	U+0906	आ )	U+0A79
इ )	U+0907	इ )	U+0A7A
ई )	U+0908	ई )	U+0A7B
उ )	U+0909	उ )	U+0A7C
ऊ )	U+090A	ऊ )	U+0A7D
ए )	U+090F	ए )	U+0A7E
ऐ )	U+0910	ऐ )	U+0A7F
ओ )	U+0913	ओ )	U+0A82
औ )	U+0914	औ )	U+0A83





## Cascading Style Sheet



- Indentation of character



### Challenges :

- Implementation of CSS standards developed by W3C regarding Indian languages
- Standards however need to be provided to those developing CSS so that by default user could have the facility to use bulleting in his own Indic languages.

### Roadmap :

Initiation of study for CSS 2.0 specifications with respect to requirements for 5 Indian languages : Hindi, Bangla, Punjabi, Kannada, Tamil



## E-Government



### Issues :

- Some applications are Completely in English.
- Some applications have static content in local language but forms in English.
- Some applications are multi-lingual but only in limited languages (e.g., English and only one local language) .

### Use cases :

#### Use Case for Land Records

- As all type of data related to land, available in these land records
- They are used for various planning processes although the manual maintenance of this land record does hinder in effective collation and analysis of the data contained in them.



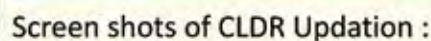


- To enable e-governance applications in Indian languages.
- Compliance with W3C standards.

## Roadmap :

- Major E-Gov Application in Indian languages need to be studied for improving access to Government through better use of the web.


Target languages : Hindi, Bangla, Marathi ,Telugu, Tamil




- CLDR HINDI

[illegible]






Common Locale Data Repository




• Some of the Screen shots of CLDR Updation :

CLDR Bengali

AM	সকাল	সকাল	সকাল	সকাল	Saturday	শনিবার	শনিবার
PM	বিকাল	বিকাল	বিকাল	বিকাল	Sunday	রবিবার	রবিবার
Sunday	রবিবার	রবিবার	রবিবার	রবিবার	Monday	সোমবার	সোমবার
Monday	সোমবার	সোমবার	সোমবার	সোমবার	Tuesday	মঙ্গলবার	মঙ্গলবার
Tuesday	মঙ্গলবার	মঙ্গলবার	মঙ্গলবার	মঙ্গলবার	Wednesday	বুধবার	বুধবার
Wednesday	বুধবার	বুধবার	বুধবার	বুধবার	Thursday	বৃহসপতি	বৃহসপতি
Thursday	বৃহসপতি	বৃহসপতি	বৃহসপতি	বৃহসপতি	Friday	শুক্রবার	শুক্রবার
Friday	শুক্রবার	শুক্রবার	শুক্রবার	শুক্রবার	Saturday	শনিবার	শনিবার
Saturday	শনিবার	শনিবার	শনিবার	শনিবার	Day	দিন	দিন
Day	দিন	দিন	দিন	দিন	Yesterday	গতকাল	গতকাল
Yesterday	গতকাল	গতকাল	গতকাল	গতকাল	The day before yesterday	গতগতকাল	গতগতকাল
The day before yesterday	গতগতকাল	গতগতকাল	গতগতকাল	গতগতকাল	Three days ago	তিন দিন আগের	তিন দিন আগের
Three days ago	তিন দিন আগের	তিন দিন আগের	তিন দিন আগের	তিন দিন আগের	Today	আজ	আজ
Today	আজ	আজ	আজ	আজ	Tomorrow	আমরকাল	আমরকাল
Tomorrow	আমরকাল	আমরকাল	আমরকাল	আমরকাল	The day after tomorrow	আমরকাল	আমরকাল
The day after tomorrow	আমরকাল	আমরকাল	আমরকাল	আমরকাল	Three days from now	তিন দিন পর	তিন দিন পর
Three days from now	তিন দিন পর	তিন দিন পর	তিন দিন পর	তিন দিন পর	Era	যুগ	যুগ
Era	যুগ	যুগ	যুগ	যুগ	Hour	ঘণ্টা	ঘণ্টা
Hour	ঘণ্টা	ঘণ্টা	ঘণ্টা	ঘণ্টা	Minute	মিনিট	মিনিট
Minute	মিনিট	মিনিট	মিনিট	মিনিট	Month	মাস	মাস
Month	মাস	মাস	মাস	মাস	Week	সপ্তাহ	সপ্তাহ
Week	সপ্তাহ	সপ্তাহ	সপ্তাহ	সপ্তাহ	Day of the Week	সপ্তাহের দিন	সপ্তাহের দিন
Day of the Week	সপ্তাহের দিন	সপ্তাহের দিন	সপ্তাহের দিন	সপ্তাহের দিন	Year	বছর	বছর
Year	বছর	বছর	বছর	বছর	Decade	দশক	দশক
Decade	দশক	দশক	দশক	দশক	Century	শতাব্দী	শতাব্দী
Century	শতাব্দী	শতাব্দী	শতাব্দী	শতাব্দী	Millennium	হাজার বছর	হাজার বছর
Millennium	হাজার বছর	হাজার বছর	হাজার বছর	হাজার বছর	Jan	জানুয়ারি	জানুয়ারি
Jan	জানুয়ারি	জানুয়ারি	জানুয়ারি	জানুয়ারি	Feb	ফেব্রুয়ারি	ফেব্রুয়ারি
Feb	ফেব্রুয়ারি	ফেব্রুয়ারি	ফেব্রুয়ারি	ফেব্রুয়ারি	Mar	মার্চ	মার্চ
Mar	মার্চ	মার্চ	মার্চ	মার্চ	Apr	এপ্রিল	এপ্রিল
Apr	এপ্রিল	এপ্রিল	এপ্রিল	এপ্রিল	May	মে	মে
May	মে	মে	মে	মে	Jun	জুন	জুন
Jun	জুন	জুন	জুন	জুন	Jul	জুলাই	জুলাই
Jul	জুলাই	জুলাই	জুলাই	জুলাই	Aug	আগস্ট	আগস্ট
Aug	আগস্ট	আগস্ট	আগস্ট	আগস্ট	Sep	সেপ্টেম্বর	সেপ্টেম্বর
Sep	সেপ্টেম্বর	সেপ্টেম্বর	সেপ্টেম্বর	সেপ্টেম্বর	Oct	অক্টোবর	অক্টোবর
Oct	অক্টোবর	অক্টোবর	অক্টোবর	অক্টোবর	Nov	নভেম্বর	নভেম্বর
Nov	নভেম্বর	নভেম্বর	নভেম্বর	নভেম্বর	Dec	ডিসেম্বর	ডিসেম্বর
Dec	ডিসেম্বর	ডিসেম্বর	ডিসেম্বর	ডিসেম্বর	Q1	১ম ষত্র	১ম ষত্র
Q1	১ম ষত্র	১ম ষত্র	১ম ষত্র	১ম ষত্র	Q2	২য় ষত্র	২য় ষত্র
Q2	২য় ষত্র	২য় ষত্র	২য় ষত্র	২য় ষত্র	Q3	৩য় ষত্র	৩য় ষত্র
Q3	৩য় ষত্র	৩য় ষত্র	৩য় ষত্র	৩য় ষত্র	Q4	৪য় ষত্র	৪য় ষত্র
Q4	৪য় ষত্র	৪য় ষত্র	৪য় ষত্র	৪য় ষত্র	1st quarter	১ম ষত্র	১ম ষত্র
1st quarter	১ম ষত্র	১ম ষত্র	১ম ষত্র	১ম ষত্র	2nd quarter	২য় ষত্র	২য় ষত্র
2nd quarter	২য় ষত্র	২য় ষত্র	২য় ষত্র	২য় ষত্র	3rd quarter	৩য় ষত্র	৩য় ষত্র
3rd quarter	৩য় ষত্র	৩য় ষত্র	৩য় ষত্র	৩য় ষত্র	4th quarter	৪য় ষত্র	৪য় ষত্র
4th quarter	৪য় ষত্র	৪য় ষত্র	৪য় ষত্র	৪য় ষত্র			



Common Locale Data Repository



• Some of the Screen shots of CLDR Updation :

CLDR Malayalam

The day before yesterday	ഇന്നിടെ ദിവസം	ഇന്നിടെ ദിവസം	ഇന്നിടെ ദിവസം	Jul	ജൂലൈ	7
Three days ago	മൂന്നു ദിവസം മുമ്പ്	മൂന്നു ദിവസം മുമ്പ്	മൂന്നു ദിവസം മുമ്പ്	July	ജൂലൈ	7
Today	ഇന്ന്	ഇന്ന്	ഇന്ന്	Aug	ഓഗസ്റ്റ്	8
Tomorrow	മുറ്റമുറ്റം	മുറ്റമുറ്റം	മുറ്റമുറ്റം	August	ഓഗസ്റ്റ്	8
The day after tomorrow	മുറ്റമുറ്റം	മുറ്റമുറ്റം	മുറ്റമുറ്റം	Sep	സെപ്റ്റംബർ	9
Three days from now	മൂന്നു ദിവസം ശേഷം	മൂന്നു ദിവസം ശേഷം	മൂന്നു ദിവസം ശേഷം	September	സെപ്റ്റംബർ	9
Era	യുഗം	യുഗം	യുഗം	Oct	ഒക്ടോബർ	10
Hour	മണിക്കൂറു	മണിക്കൂറു	മണിക്കൂറു	October	ഒക്ടോബർ	10
Minute	മിനിറ്റു	മിനിറ്റു	മിനിറ്റു	Nov	നവംബർ	11
Month	മാസം	മാസം	മാസം	November	നവംബർ	11
Week	ആഴ്ച	ആഴ്ച	ആഴ്ച	Dec	ഡിസംബർ	12
Day of the Week	ആഴ്ചയിലെ ദിവസം	ആഴ്ചയിലെ ദിവസം	ആഴ്ചയിലെ ദിവസം	December	ഡിസംബർ	12
Year	വർഷം	വർഷം	വർഷം	Q1	1st quarter	1st quarter
Decade	ദശകം	ദശകം	ദശകം	Q2	2nd quarter	2nd quarter
Century	ശതകം	ശതകം	ശതകം	Q3	3rd quarter	3rd quarter
Millennium	ഹাজার വർഷം	ഹাজার വർഷം	ഹাজার വർഷം	Q4	4th quarter	4th quarter
Jan	ജനുവരി	ജനുവരി	ജനുവരി	1st quarter	1st quarter	1st quarter
Feb	ഫെബ്രുവരി	ഫെബ്രുവരി	ഫെബ്രുവരി	2nd quarter	2nd quarter	2nd quarter
Mar	മാർച്ച്	മാർച്ച്	മാർച്ച്	3rd quarter	3rd quarter	3rd quarter
Apr	ഏപ്രിൽ	ഏപ്രിൽ	ഏപ്രിൽ	4th quarter	4th quarter	4th quarter
May	മേ	മേ	മേ			
Jun	ജൂൺ	ജൂൺ	ജൂൺ			
Jul	ജൂലൈ	ജൂലൈ	ജൂലൈ			

CLDR Assamese

Jul	জুলাই	7
July	জুলাই	7
Aug	আগষ্ট	8
August	আগষ্ট	8
Sep	সেপ্টেম্বর	9
September	সেপ্টেম্বর	9
Oct	অক্টোবর	10
October	অক্টোবর	10
Nov	নভেম্বর	11
November	নভেম্বর	11
Dec	ডিসেম্বর	12
December	ডিসেম্বর	12
Q1	১ম ষত্র	1st quarter
Q2	২য় ষত্র	2nd quarter
Q3	৩য় ষত্র	3rd quarter
Q4	৪য় ষত্র	4th quarter
1st quarter	১ম ষত্র	1st quarter
2nd quarter	২য় ষত্র	2nd quarter
3rd quarter	৩য় ষত্র	3rd quarter
4th quarter	৪য় ষত্র	4th quarter





## Common Locale Data Repository



### DRAFT LOCALE DATA for HINDI :

#### भारतीय कैलेंडर (Indian Calendar)

##### 1. विक्रम संवत्

##### महीना/मास (Months)

चैत्र  
वैशाख  
ज्येष्ठ  
आषाढ़  
श्रावण  
भाद्र  
आश्विन  
कार्तिक  
अग्रहायण  
पौष  
माघ  
फाल्गुन

#### Collation sequence sorting

अ आ इ ई उ ऊ ऋ ए औ औ  
क ख ग घ ङ  
च छ झ ञ  
ट ठ ड ढ ण  
त थ द ध न  
प फ ब भ म  
य र ल व  
श ष स ह  
उ ण  
क्ष व ज ञ  
ॐ  
ॐ  
ॐ  
ॐ  
ख ङ फ  
१ २ ३ ४ ५ ६ ७ ८ ९ ०

#### समय अवधि (days)

आज  
कल  
परसो  
नरसो

#### दिन (Time Period of A Day)

रात / सुबह / प्रभात  
पूर्वाह्न  
अपराह्न  
मध्याह्न / दोपहर  
संध्या / शाम / सांझ  
रात / रात्रि

#### तिमाही (Quarter)

दूसरी तिमाही / छमाही  
तीसरी तिमाही  
वर्ष / साल  
दशवर्षी  
शतावर्षी  
सहस्रवर्षी



## Common Locale Data Repository



### Draft locale data for Bengali :

S/No	বাংলা মাস (Bangla Months)
1	বৈশাখ
2	জ্যৈষ্ঠ
3	আষাঢ়
4	শ্রাবণ
5	ভাদ্র
6	আশ্বিন
7	কার্তিক
8	অগ্রহায়ণ
9	পৌষ
10	মাঘ
11	ফাল্গুন
12	চৈত্র

S/No	বাংলা সপ্তাহের বার (Bangla Week's day)
1	রবিবার
2	সোমবার
3	মঙ্গলবার
4	বুধবার
5	বৃহস্পতিবার / লক্ষ্মীবার
6	শুক্রবার

S/No	English Name	Bangla Name
1	Era	যুগ
2	Hour	ঘণ্টা
3	Minute	মিনিট
4	Second	সেকেন্ড
5	Month	মাস
6	Week	সপ্তাহ
7	Year	বছর
8	Zone	এলাকা
9	BC	খ্রিস্টপূর্ব
10	AD	খ্রিস্টাব্দ
11	Day	দিন









## WEB ACCESSIBILITY INITIATIVE



### Challenges :

- Make Web content accessible to people with disabilities w.r.t Indian languages
- WCAG 2.0 Guidelines for success criteria vis-a-vis selected recommendations relevant to Indian context

### Initiative in India :

- "Guidelines for Indian Government websites" by NIC , Govt. of India
- STQC Implementing WCAG 2.0 Accessibility through Website Quality Certification
- Centre for Internet and Society developing authorized translation of WCAG 2.0 Guidelines

### Roadmap :

- Meet WCAG 2.0 guidelines & techniques w.r.t Indian languages
- Initiation with Hindi, Bangla, Marathi, Telugu, Tamil



## Internet Domain Names (IDN's)



### Issues

- Spoofing issues- Homographs
  - Characters looks similar in address bar  
eg. 1. क & फ  
कमल फमल
  - No two scripts should get mixed
  - Normal generic rules have to be there with some added restrictions as per language demands are required
- Spelling Variants  
eg. "हिंदी" and "हिन्दी"





## Internet Domain Names (IDN's)



- **Browsers related Issues**
  - No backward compatibility
  - Conversion from Unicode to Punycode is available in IE7 and onwards
  - Firefox directly converts Unicode to Punycode
  - Some rendering issues in different browsers for different Indian languages
- **Vertical conjuncts**

Tamil	இந்தியா
Telugu	ఇండియా

DN Draft policy for Indian Language



## Unified Language Character Table



**Malayalam Language Table**

Unicode	Character	Description
<b>Various signs</b>		
0D02	ഓ	MALAYALAM SIGN ANUSVARAM
0D03	ഔ	MALAYALAM SIGN VISARGAM
<b>Independent Vowels</b>		
0D05	അ	MALAYALAM LETTER A
0D06	ആ	MALAYALAM LETTER AA
0D07	ഇ	MALAYALAM LETTER I
0D08	ഈ	MALAYALAM LETTER II
0D09	ഉ	MALAYALAM LETTER U
0D0A	ഊ	MALAYALAM LETTER UU
0D0B	എ	MALAYALAM LETTER VOCALIC E
0D0C	ഐ	MALAYALAM LETTER EE
0D10	ഐ	MALAYALAM LETTER AI
0D12	ഒ	MALAYALAM LETTER O
0D13	ഓ	MALAYALAM LETTER OO
0D14	ഔ	MALAYALAM LETTER AU
0D15	ക	MALAYALAM LETTER KA
0D16	ഖ	MALAYALAM LETTER KHA
0D17	ഗ	MALAYALAM LETTER GA
0D18	ഘ	MALAYALAM LETTER GHA
0D19	ങ	MALAYALAM LETTER NGA
0D1A	ച	MALAYALAM LETTER CA
0D1B	ച	MALAYALAM LETTER CHA
0D1C	ട	MALAYALAM LETTER JA
0D1D	ഠ	MALAYALAM LETTER BHA
0D1E	ണ	MALAYALAM LETTER NYA

**Tamil Language Table**

Unicode	Character	Description
0B83	ஃ	TAMIL SIGN VISARGA = aytham
<b>Independent vowels</b>		
0B85	அ	TAMIL LETTER A
0B86	ஆ	TAMIL LETTER AA
0B87	இ	TAMIL LETTER I
0B88	ஈ	TAMIL LETTER II
0B89	உ	TAMIL LETTER U
0B8A	ஊ	TAMIL LETTER UU
0B8E	ஐ	TAMIL LETTER E
0B8F	ஐ	TAMIL LETTER EE
0B90	ஔ	TAMIL LETTER AI
0B92	ஓ	TAMIL LETTER O
0B93	ஔ	TAMIL LETTER OO
0B94	ஔ	TAMIL LETTER AU
<b>Consonants</b>		
0B95	க	TAMIL LETTER KA
0B99	ங	TAMIL LETTER NGA
0B9A	ச	TAMIL LETTER CA
0B9C	ஜ	TAMIL LETTER JA
0B9E	ந	TAMIL LETTER NYA
0B9F	ல	TAMIL LETTER LA
0BA3	ழ	TAMIL LETTER NNA
0BA4	ழ	TAMIL LETTER TA









## Script Grammar



### Script Grammar – Marathi

#### 8.1.1. CONSONANT SET: VALID / INVALID

Basic Consonants arranged as per their vargas (Please propose change in shape, if any)

क	ख	ग	घ	ङ
च	छ	ज	झ	ञ
ट	ठ	ड	ढ	ण
त	थ	द	ध	न
प	फ	ब	भ	म
य	र	ल	व	श
स	ह	ळ		

Ligatures: क्षत्रज्ञ are not listed

Nukta Consonants INVALID

#### VOWEL SET:

(Please propose change in shape, if any)

अ आ इ ई उ ऊ ऋ ॠ ए ऐ ओ औ

#### MATRA SET

(Please propose change in shape, if any)

ा ि ी ु ॄ ॅ ष ॆ े

ॆ is rarely used in Marathi - There is only one word i.e. - कसृष्टी, कसृष्टि.



## Challenges



- Multiplicity of Languages
- Evolution of Orthography
- Lack of Standardization
- Enabling Mobile web in Indian languages
- Initiation of study for a W3C recommendations with respect to requirements for Indian languages
- Adoption of W3C standards in terms of Internationalization
- Indian Websites should be fully W3C Complaint
- E-Gov Application in Indian languages need to be studied for better use of the web.





## Integrated challenges for W3C India Office



- **Language Tag**
  - Initiative in vetting / modification / developing Language Tags in all 22 official Indian languages as well as regional dialectical variation of Indian languages.
- **CLDR**
  - Six Languages in CLDR Hindi , Nepali, Bengali , Assamese, Malayalam and Gujarati are finalized. Other languages are in process
- **Revised Inscript Keyboard Layout** – Enhanced to incorporate additional characters as per Unicode 5.1. C-DAC, IBM, Microsoft & Redhat involved in this initiative.



Thank You



## Indian Language Phonemes and Creation of Pronunciation Lexicon in W3C Framework

**Dr. Shyamal Kumar Das Mandal**

[shyamal.dasmandal@kolkatacdac.in](mailto:shyamal.dasmandal@kolkatacdac.in)

Centre for Development of Advanced Computing (C-DAC), Kolkata

[www.cdackolkata.in](http://www.cdackolkata.in)

*6<sup>th</sup> May 2010*

*World Wide Web: Technology, Standards and Internationalization*

What is Pronunciation Lexicon?

Representation of Pronunciation information of the  
Lexicon items along with its Grapheme information

Why Pronunciation Lexicon ?

It required for the development of Speech  
technology such as Text to Speech Synthesis and  
Automatic Speech Recognition



## Multiple pronunciations for the same orthography



### Problem no.1 → Homographs

Homographs means there are words with the same spelling and different meanings but different pronunciations.

```
<?xml version="1.0" encoding="UTF-8"?>
<lexicon version="1.0"
xmlns="http://www.w3.org/2005/01/pronunciation-lexicon"
alphabet="ipa" xml:lang="bn">
  <lexeme>
    <grapheme>সরল</grapheme>
    <phoneme>ʃɔrɔl</phoneme>
    <!-- IPA string is: "ʃɔrɔl" -->
  </lexeme>
  <lexeme>
    <grapheme>সরল</grapheme>
    <phoneme>ʃɔrlo</phoneme>
    <!-- IPA string is: "ʃɔrlo" -->
  </lexeme>
</lexicon>
```

### Solution under the existing PLS specification

- ❖ Using "Role" attribute under the Lexeme element
- ❖ Using prefer attribute under the Phoneme element

## Proposed Solution



### Solution 1:

The "pos" can be an optional attribute under the phoneme element which indicates the detail information for obtaining the pronunciation

```
<?xml version="1.0" encoding="UTF-8"?>
<lexicon version="1.0" xmlns="http://www.w3.org/2005/01/pronunciation-lexicon"
alphabet="ipa" xml:lang="bn">
  <lexeme>
    <grapheme>সরল</grapheme>
    <phoneme pos="adjective">ʃɔrɔl</phoneme>
    <!-- IPA string is: "ʃɔrɔl" -->
    <!-- Itrans is: "sarala" -->
    <!-- Meaning is: "easy" -->
    <phoneme pos="verb">ʃɔrlo</phoneme>
    <!-- IPA string is: "ʃɔrlo" -->
    <!-- Itrans is: "sarala" -->
    <!-- Meaning is: "moved" -->
    <phoneme pos="null">ʃɔrɔl</phoneme>
    <!-- IPA string is: "ʃɔrɔl" -->
    <!-- Itrans is: "sarala" -->
    <!-- Meaning is: "easy" -->
  </lexeme>
</lexicon>
```

#### Advantage:

- ❖ Reduction of lexeme numbers.
- ❖ Removal of disambiguity



## Proposed Solution



### Solution 2:

The <lexeme> element may contain optionally one or more <pos> element.

```
<?xml version="1.0" encoding="UTF-8"?>
<lexicon version="1.0" xmlns="http://www.w3.org/2005/01/pronunciation-lexicon"
  alphabet="ipa" xml:lang="bn">
  <lexeme>
    <grapheme>সরল</grapheme>
    <pos> adjective </pos>
    <phoneme> /sɔɾol/ </phoneme>
    <!-- IPA string is: " /sɔɾol" -->
    <!-- Itrans is: "sarala" -->
    <!-- Meaning is : "easy" -->
    <pos> verb </pos>
    <phoneme> /sɔɾla/ </phoneme>
    <!-- IPA string is: " /sɔɾla" -->
    <!-- Itrans is: "sarala" -->
    <!-- Meaning is: "moved" -->
    <pos> null </pos>
    <phoneme> /sɔɾol/ </phoneme>
    <!-- IPA string is: " /sɔɾol" -->
    <!-- Itrans is: "sarala" -->
    <!-- Meaning is: "easy" -->
  </lexeme>
</lexicon>
```

#### Advantage:

- ❖ Reduction of lexeme numbers.
- ❖ Removal of disambiguity



### Problem no. 2 :

Ideolectal variation of the same orthographic information

## Proposed Solution

```
<?xml version="1.0" encoding="UTF-8"?>
<lexicon version="1.0" xmlns="http://www.w3.org/2005/01/pronunciation-lexicon"
  alphabet="ipa" xml:lang="bn">
  <lexeme>
    <grapheme>উন্ট্রিফ</grapheme>
    <phoneme prefer="true">untriɸ</phoneme>
    <!-- IPA string is: "untriɸ" -->
    <phoneme>unotiriɸ</phoneme>
    <!-- IPA string is: "unotiriɸ" -->
  </lexeme>
</lexicon>
```



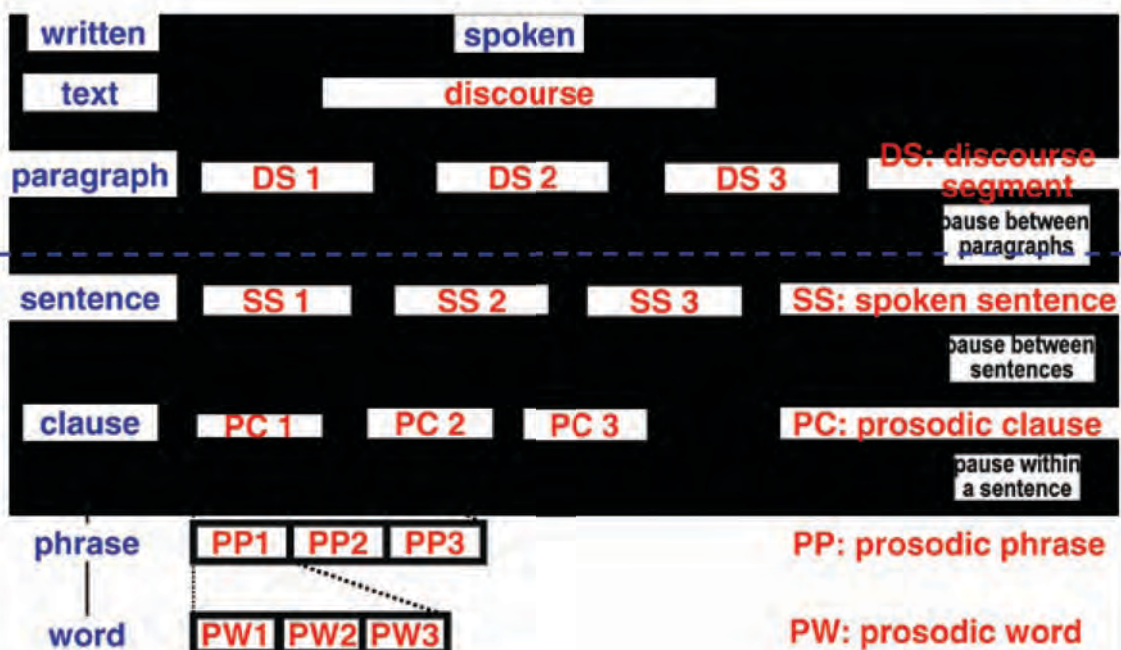
## Need of morphological information



In some of the languages like Bengali not only POS information but also morphological information especially in case of verb finiteness and honorificity information are very crucial in determining the pronunciation of a homograph

```
<?xml version="1.0" encoding="UTF-8"?>
<lexicon version="1.0" xmlns="http://www.w3.org/2005/01/pronunciation-lexicon"
  alphabet="ipa" xml:lang="bn">
  <lexeme>
    <grapheme>করে</grapheme>
    <phoneme: pos= "VM.3.prs.smp.dcl.fin.n.n.n" >kɔre</phoneme>
    <!-- IPA string is: "kɔre" -->
    <!-- Itrans is: "kare" -->
    <!-- Meaning is: "do/does" -->
    <phoneme: pos= "VM.0.0.0.0.nfn.n.n.n" >kore</phoneme>
    <!-- IPA string is: "kore" -->
    <!-- Itrans is: "kare" -->
    <!-- Meaning is: "having done" -->
  </lexeme>
</lexicon>
```

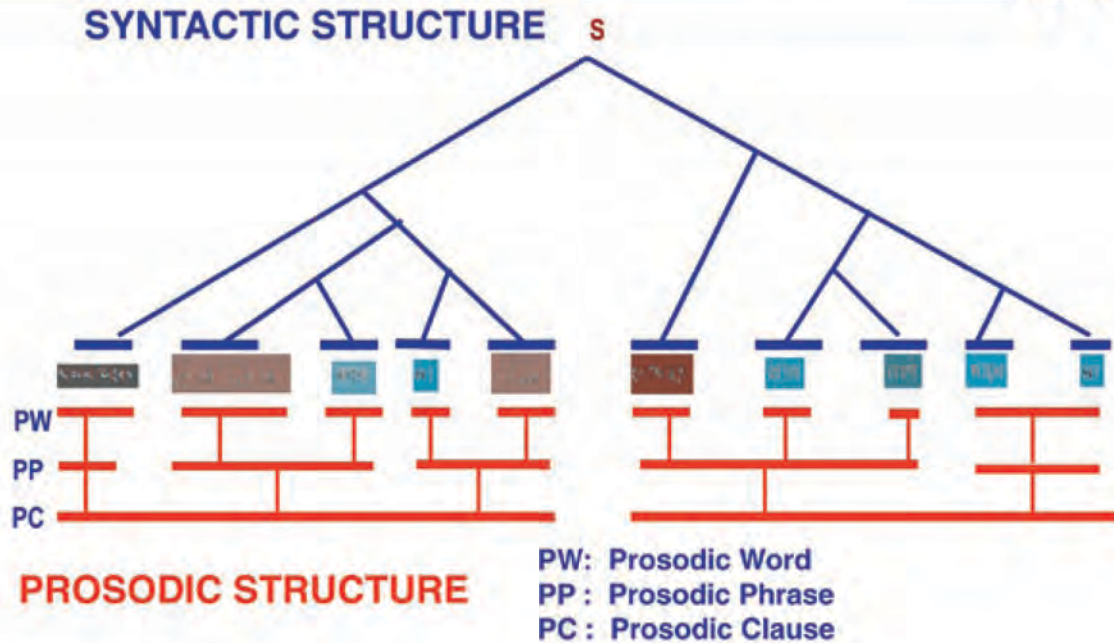
## Prosodic Structure





An Example of Disparity Between the Syntactic Structure and the Prosodic Structure

सी डैक  
CDAC



© Hiroya Fujisaki

Pitch contour

सी डैक  
CDAC

```
<?xml version="1.0"?> <speak version="1.1"
xmlns="http://www.w3.org/2001/10/synthesis"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.w3.org/2001/10/synthesis
http://www.w3.org/TR/speech-synthesis11/synthesis.xsd"
xml:lang="en-US">

<prosody contour="(0%,+20Hz) (10%,+30%) (40%,+10Hz)">
good morning
</prosody>
</speak>
```



## Emphasis Element



```
<?xml version="1.0"?> <speak version="1.1"
xmlns="http://www.w3.org/2001/10/synthesis"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.w3.org/2001/10/synthesis
http://www.w3.org/TR/speech-synthesis11/synthesis.xsd"
xml:lang="en-US">
```

That is a <emphasis> big </emphasis> car!

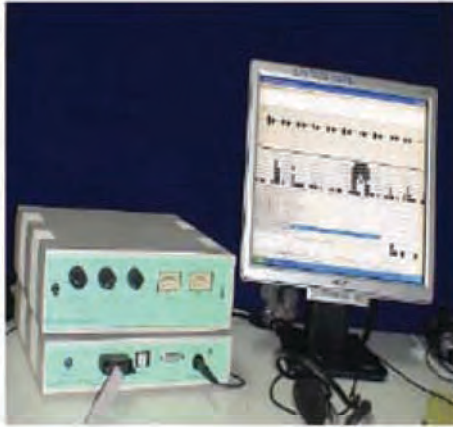
That is a <emphasis level="strong"> huge </emphasis>  
bank account! </speak>



धन्यवाद



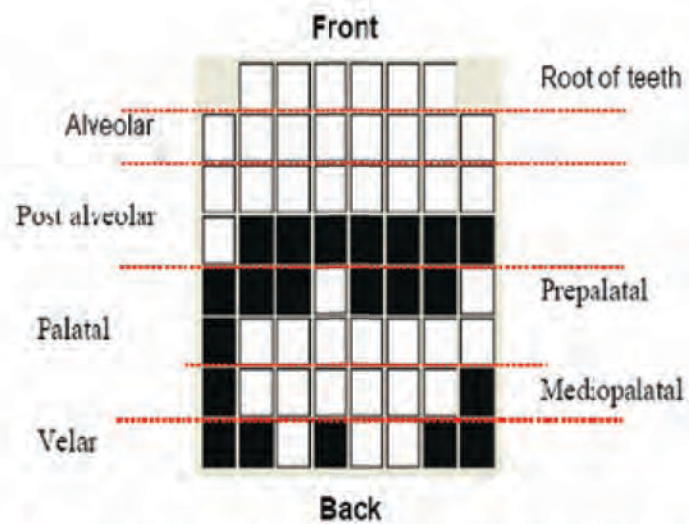




## WinEPG System



## Custom-made Palates



### Articulatory regions of the custom made palate



Plosive/stop three variables were analyzed

- Place of closure
- Place of release,
- Amount of contact area for both the cases

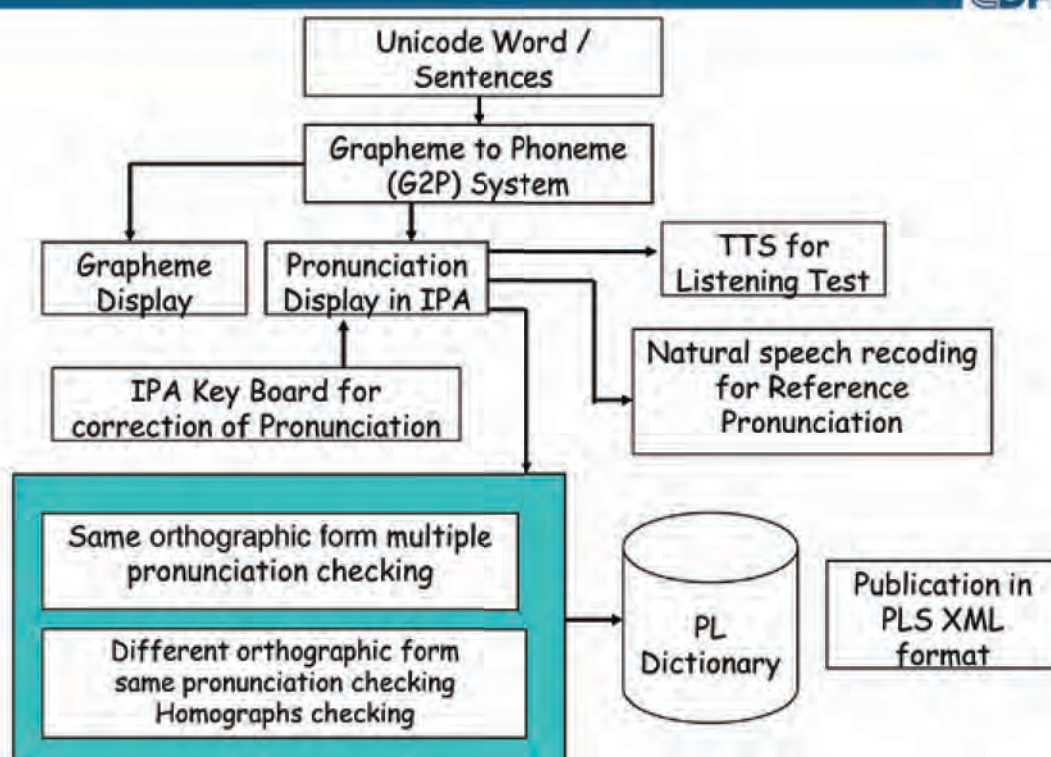
#### a) Place of closure

The beginning of the occlusion was identified when contact point spread from left to right end uninterruptedly showing a complete separation of the front and back halves of the palate



Closure

### Functional Block diagram of the IDE





## Features of the above developed IDE for PLS creation in Indian languages



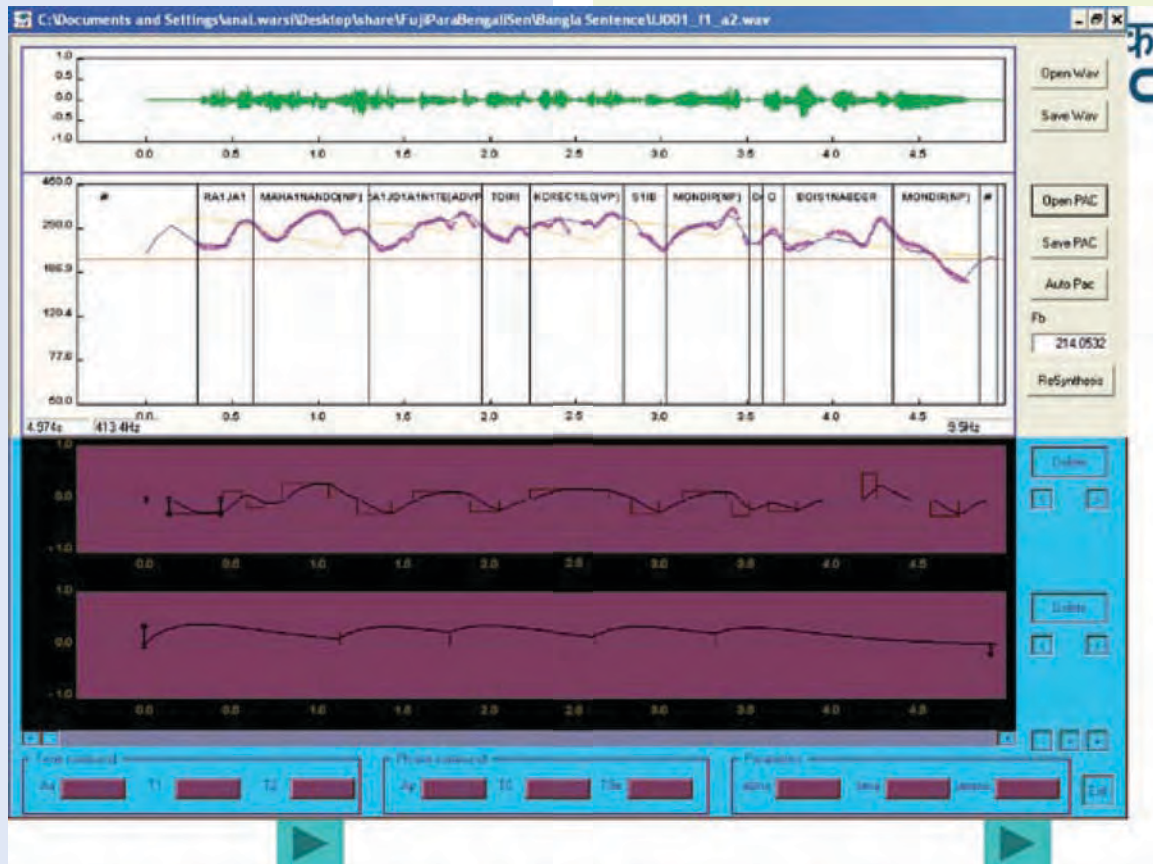
- ❑ System will include the G2P conversion system for Bangla and have the facility to include new G2P conversion system for other languages.
- ❑ System will be developed for both the platforms (Windows and Linux)
- ❑ System will Provide the facility for recording of reference pronunciation and Listening facility using available TTS system
- ❑ System will take Unicode input either in from of word or sentences and generate the pronunciation of each word.
- ❑ The PL will be store as per W3C PLS standards

..  
..



```
<?xml version="1.0" encoding="UTF-8" ?>
<!-- This pronunciation lexicon is licensed under the GPL. -->
<lexicon version="1.0" xmlns="http://www.w3.org/2005/01/pronunciation-lexicon"
  xmlns:xsl="http://www.w3.org/2001/XMLSchema-instance"
  xsl:schemaLocation="http://www.w3.org/2005/01/pronunciation-lexicon
    http://www.w3.org/TR/2007/CR-pronunciation-lexicon-20071212/pls.xsd" alphabet="ipa"
  xml:lang="de">
  <lexeme>
    <grapheme>केय</grapheme>
    <phoneme>keu</phoneme>
  </lexeme>
  <lexeme>
    <grapheme>केर</grapheme>
    <phoneme>ker</phoneme>
  </lexeme>
  <lexeme>
    <grapheme>केज</grapheme>
    <phoneme>kej</phoneme>
  </lexeme>
  <lexeme>
    <grapheme>केत</grapheme>
    <phoneme>ket</phoneme>
  </lexeme>
  </lexicon>
```





धन्यवाद





## The Creation of PLS in Indian languages



- ❑ Standardization of Phonetic Representation of Indian language Phonemes
- ❑ Creation of PLS in Indian Languages

## Present Pronunciation Lexicon Markup Language Definition



Elements	Attributes	Description
<u>&lt;lexicon&gt;</u>	version xml:base xmlns xml:lang alphabet	root element for PLS
<u>&lt;meta&gt;</u>	name http-equiv content	meta data container element
<u>&lt;metadata&gt;</u>		meta data container element
<u>&lt;lexeme&gt;</u>	xml:id	the container element for a single lexical entry
<u>&lt;grapheme&gt;</u>	orthography	contains orthographic information for a lexeme
<u>&lt;phoneme&gt;</u>	prefer alphabet	contains pronunciation information for a lexeme
<u>&lt;alias&gt;</u>	prefer	contains acronym expansions and words' substitutions
<u>&lt;example&gt;</u>		contains an example of the usage for a lexeme



## What is PLS of W3C?

The Pronunciation Lexicon Specification (PLS) is designed to enable interoperable specification of pronunciation information for both [ASR](#) and [TTS](#) engines within voice browsing applications

## How it is used in TTS and ASR?

The PLS is the standard format of the documents referenced by the [<lexicon>](#) element of [SSML](#). The PLS engine will load the external PLS document and transparently apply the pronunciations during the processing of the [SSML](#) document. An application may contain several distinct PLS documents to be used in different points of the application.

If a [pronunciation lexicon](#) is referenced by a [SRGS](#) grammar it can allow multiple pronunciations of the word in the grammar to accommodate different speaking styles

## Multiple orthographies with same pronunciation

### Homophones

Homophones means words with different spellings and different meanings but the same pronunciation

```
<?xml version="1.0" encoding="UTF-8"?>
<lexicon version="1.0"
  xmlns="http://www.w3.org/2005/01/pronunciation-lexicon"
  alphabet="ipa" xml:lang="bn">
  <lexeme>
    <grapheme>কুল</grapheme>
    <phoneme>kul</phoneme>
  </lexeme>
  <lexeme>
    <grapheme>কুল</grapheme>
    <phoneme>kul</phoneme>
  </lexeme>
</lexicon>
```



## Multiple orthographies with same pronunciation



```
<?xml version="1.0" encoding="UTF-8"?>
<lexicon version="1.0"
  xmlns="http://www.w3.org/2005/01/pronunciation-lexicon"
    alphabet="ipa" xml:lang="bn">
  <!-- English entry showing how alternative spellings are
  handled -->
  <lexeme>
    <grapheme>উনিশ</grapheme>
    <grapheme>উনিশ</grapheme>
    <phoneme>uniʃ</phoneme>
    <!-- IPA string is: "uniʃ" -->
  </lexeme>
</lexicon>
```



### What is Pronunciation Lexicon?

Representation of Pronunciation information of the Lexicon items along with its Grapheme information

### Why Pronunciation Lexicon ?

It required for the development of Speech technology such as Text to Speech Synthesis and Automatic Speech Recognition



## Mobile Web Accessibility

Two fantastic developments and one crying need !!

Ajay Kolhatkar, PhD  
Research Analyst,  
Web 2.0 Research Lab, SETLabs,  
Infosys Technologies Limited  
Ajay\_Kolhatkar@infosys.com

## MOBILE PHONES - MAN'S BEST INVENTION YET?

Well..maybe the third best !



## Mobile Phones – Just short of multi sensory

- See
- Hear
- Touch
- Speak
- Sense
  
- Smell
- Taste

## Mobile Phones – Just short of multi sensory

- See – Inbuilt cameras, sometimes one too many
- Hear – Speech input, Voice activated commands
- Touch – Vibration, Touch Screen,
- Speak – Speakers, Earphones
- Sense
  - Motion – Accelerometers, Compass
  - Location – GPS, Cell Tower Proximity, Triangulation
  - Navigation – Maps , Routes
  
- Smell – Who knows... may be coming soon
- Taste – Nah... I don't think so



## WORLD WIDE WEB

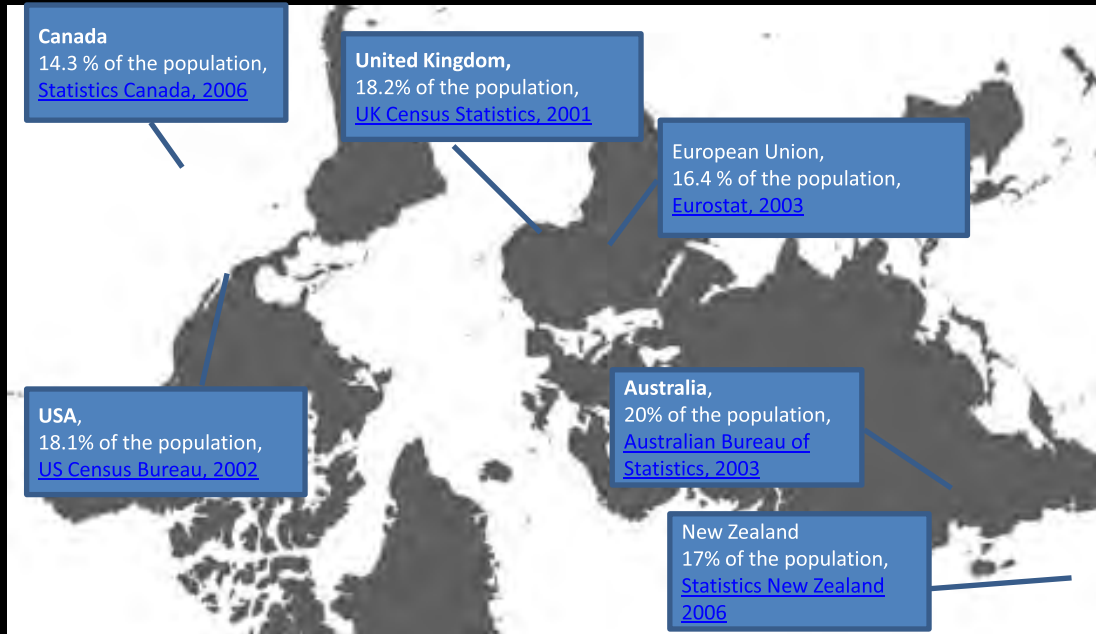
A truly global phenomenon

## ACCESSIBILITY (FOR DIFFERENTLY ABLE, AND ELDERS)

Long way to go yet !



## Disability Statistics



## Demographics of Disability

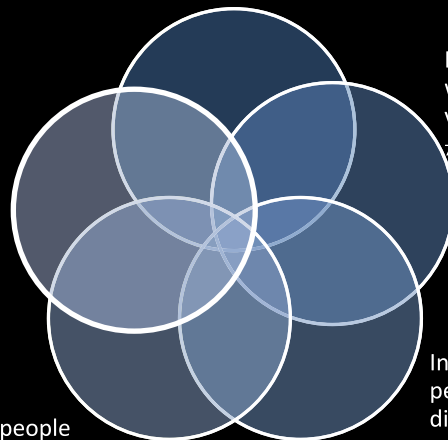
Worldwide, 650 million people have a disability

- three out of every 10 families are touched by a disability

### By 2015

- 20% of the EU will be over 65 years of age
- the number of people aged 60 or over will double in the next 30 years
- the number aged 80 or over will increase by 10% by 2050 .

Across European Union approximately 46 million people have a disability



In 2001, 180 million people worldwide were blind or visually impaired

7.7 million people in the United States

In the United States, one in five people have some kind of disability and one in 10 has a severe disability

- That's approximately 54 million Americans



## Economics of Disability

- The UK Government estimates
  - their combined spending power is in excess of £200 bn broken down into the following:
  - 8.5 million disabled people with a combined spending power of £40 bn;
  - 20 million people over the age of 50 with a combined spending power of £160 bn
- In USA, the discretionary income of people with disabilities is \$175 billion!

## Some Insights from Differently Able Users

- The Web plays an important role and has significant benefits for people with disabilities
  - Of the 54 million Americans with a disability, 4 in 10 are online
  - These users spend more time logged on and surfing the Internet than nondisabled users
  - On average, they spend 20 hours per week online
- According to the Harris Poll, 48 percent of respondents with disabilities reported that the quality of their lives had been significantly improved by the Internet compared to 27 percent of respondents without a disability



## Status of Web Accessibility

- About 97% of websites fail to meet the most basic requirements for accessibility
- The number of people with disabilities – and income to spend – is likely to increase
  - The likelihood of having a disability increases with age, and the overall population is aging
- Only 23% US federal government websites, 11% Non Profit Organization's websites and 6% corporate websites in the US are accessible

## Legislations and Regulations

United States of America	Rehabilitation Act, <a href="#">Section 504</a> (USA, 1973) Americans with Disabilities Act ( <a href="#">ADA</a> ) (USA, 1990) – Title II & Title III Amended <a href="#">Section 255</a> of the Communications Act (USA, 1996) Rehabilitation Act Amendment, <a href="#">Section 508</a> (USA, 1998)
United Kingdom	Disability Discrimination <a href="#">Act</a> of 1995 (UK, 1995) Disability Rights Commission (DRC) published a Code of Practice for "Rights of Access – Goods, Facilities, Services and Premises" (UK, 2002) DRC Published Code of Practice for Website Accessibility ( <a href="#">PAS78</a> ) (UK, 2006)
Australia	Disability Discrimination <a href="#">Act</a> (1992) WWW Accessibility (Disability) <a href="#">Policy</a> (Australia, 2000)
Canada	Canadian Human Rights <a href="#">Act</a> (Canada, 1977) Employment Equity Act (Canada, 1995) Ontarians with Disabilities <a href="#">Act</a> (Ontario, Canada, 2001) Common Look and Feel <a href="#">Standards for the Internet</a> (Canada, 2006)
Germany	<a href="#">Ordinance</a> on Barrier Free Information Technology or BITV (Germany, 2002)
France	Equal Opportunities <a href="#">Rights</a> (France, 2004)
Netherlands	Dutch <a href="#">Law</a> on Quality of Government Websites (2006)
European Union	Unified Web Evaluation <a href="#">Methodology</a> 1.0 (2006)



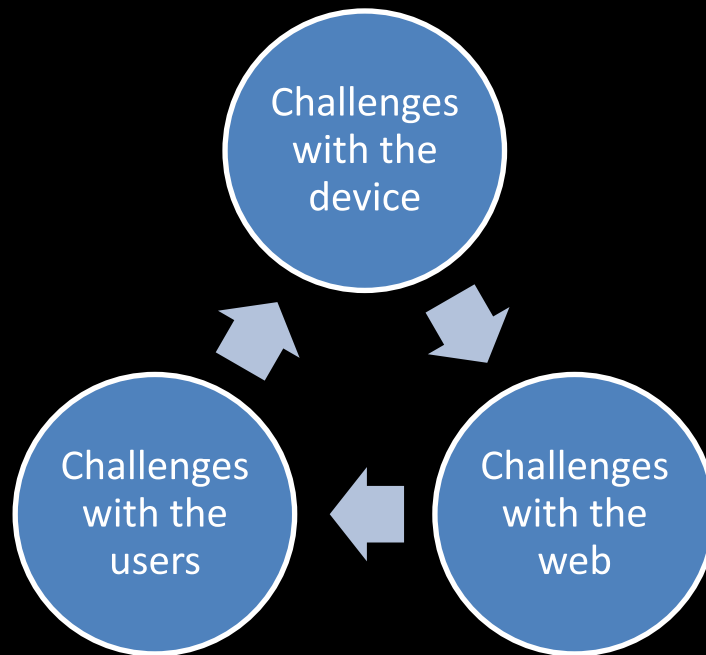
## Web Guidelines

- World Wide Web Consortia's (W3C's) Web Accessibility Initiative (WAI).
- Web Content Accessibility Guidelines (WCAG) 1.0 – released in 1999
- WCAG 2.0 – released in 2008

## MOBILE WEB ACCESSIBILITY?



## Mobile Web Accessibility



## Now add the complexity of Indian Languages

- Scripts
- Nuances
- Spoken – Written language divide
- Hinglish
- Tamilish and many others ...lishes



## Hopes !!

- Guidelines – MWBP by W3C
- Standards – HTML 5
- Product Companies
  - Nokia Symbian OS – supports 3<sup>rd</sup> Party Apps
  - Apple iPhone 3G s (VoiceOver & Zoom)
  - Google Android (TalkBack, Emacspeak)
  - RIM Blackberry
  - LG, Samsung, Sony Ericsson, Motorola
- Software
  - MobileSpeak (codefactory)
  - Oratio for Blackberry(codefactory)
  - TALKS (Nuance)
  - TalkBack (Google)
  - VoiceMode – Speech to Text for Mobile
  - KNFB

???

Tested you patience didn't I?





## High performance data intensive computing related issues

Vipin Chaudhary  
Computational Research Laboratories  
[vipin@crlindia.com](mailto:vipin@crlindia.com)  
[www.crlindia.com](http://www.crlindia.com)

1/13/2011

W3C 2010



## Semantic Web Benefits Widespread

- **Sensors widespread**
  - Too much data and not enough knowledge
  - **Semantic Sensor Web**
    - Leverage Open Geospatial Consortium and W3C
    - Annotate sensor data with spatial, temporal and thematic semantic metadata
    - **Web centric information infrastructure**
      - Collect, model, store, retrieve, share, manipulate, analyze, visualize information about sensors and sensor observations of phenomena.
- **Weather Forecast**
  - **Analyze and forecast events**
    - rainfall, drought, blizzard, cyclone, ...
- **Epidimeology**
  - **Analyze spread of diseases**



1/13/2011

W3C 2010





## Security data

- Video images
- Phone logs & content
- Email trails & content
- Financial transactions
- Travel logs
- Social networks (real and online)



## Terrorism Scenario

- Terrorist group contacts an explosive expert
  - Get a list of items he needs.
  - They then contact a set of people to purchase these items separately and deliver them to the expert.
- An ontological inference engine along with appropriate semantic data can be used to correlate
  - phone data
  - data base of people and their profession
  - knowledge of materials that make an explosive device
  - purchase logs of various materials
  - properties of materials
  - list of groups and members with disruptive intent
- can together point out a potential threat





## Scenario: Terror Group plotting



p4



p10

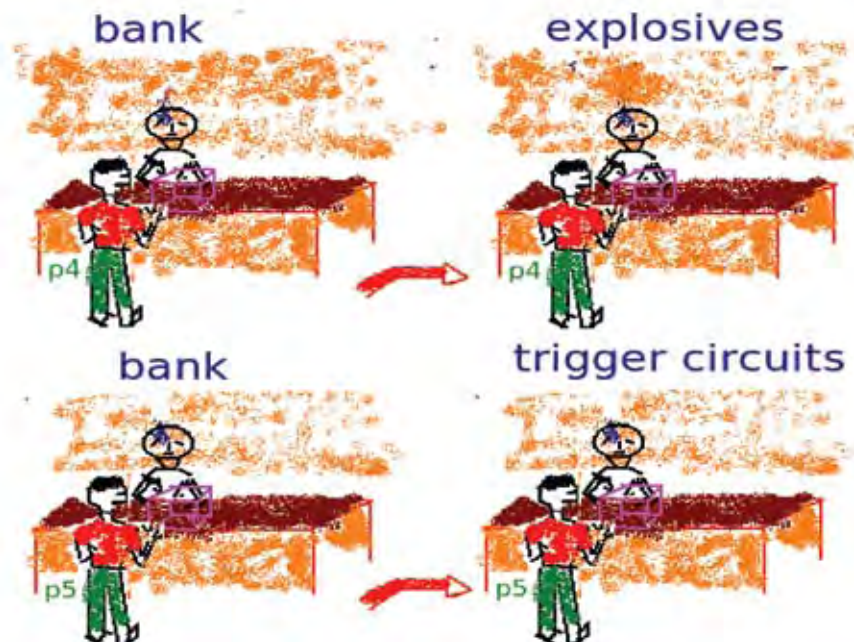


p5



p11

## Scenario: Transactions





## Scenario: p4,p5 phoning explosive expert



p4



p9



or

detection



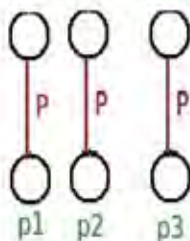
p5



p9

## Ontological relationships for inferring security threat

milk camera battery

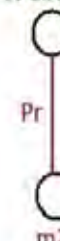


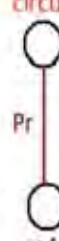
P :-purchase records of purchases from stores


explosive  
material


Pr :- property

tv circuit


energy  
storage

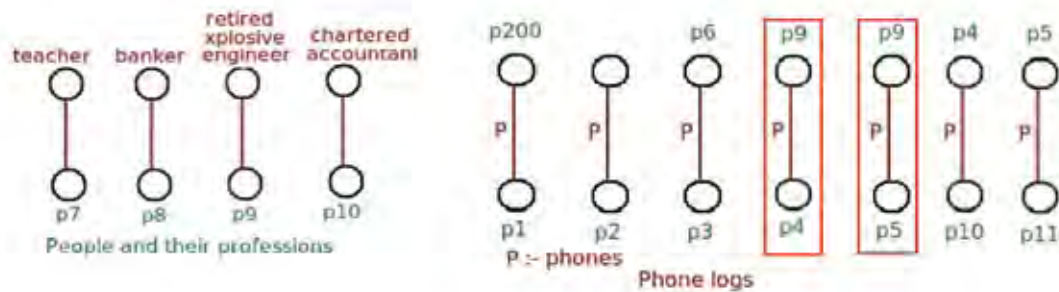
trigger  
circuit


- Ontological records of purchases and properties of the materials
- The items in red indicate potential threats as items with property explosive and trigger are purchased





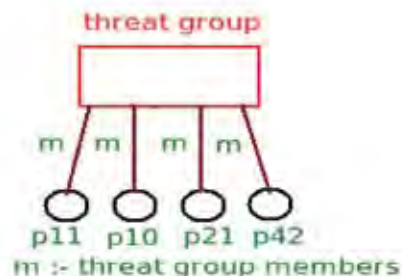
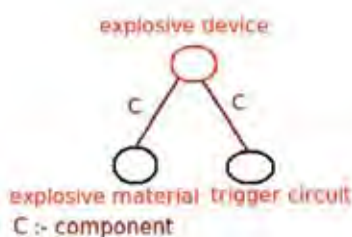
## Ontological relationships for inferring security threats



- Ontological information about people and their professions
- Database of phone logs between peoples. The item in red box shows contact with an explosive expert.



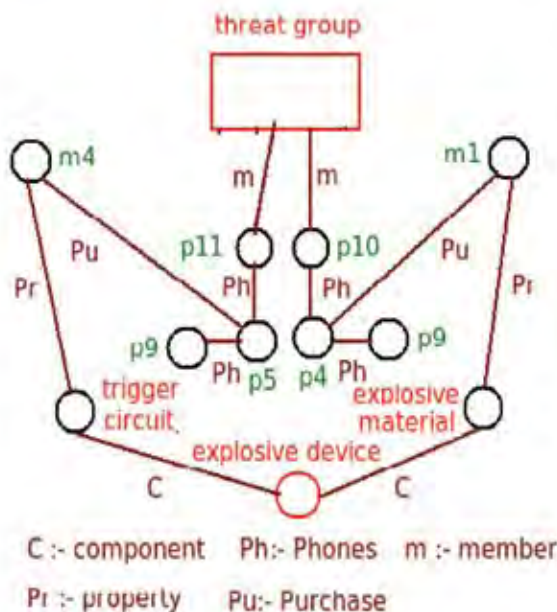
## Ontological relationships for inferring security threats



- Component structure of complex items like explosive device
- Data base of people belonging to potential threat groups



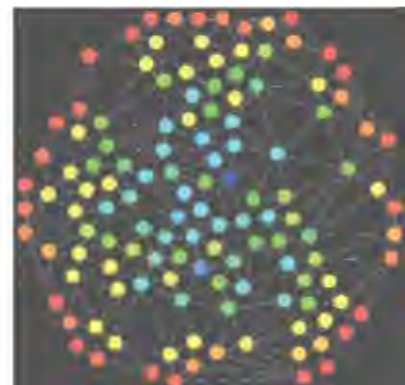
## Linking all ontological relationships to identify security threats



- Threat group members **p10** and **p11**, are linked by phone logs to **p4** and **p5**
  - who purchase **m1** and **m4** which are explosive materials and trigger circuit that make an explosive device.
- Additionally **p4** and **p5** are linked by phone logs to **p9**
  - who is an retired explosive device expert.
- Inference from a variety of semantic logs one can identify a potential security threat

## Realistic Semantic Graph Analysis is Data Intensive and needs HPC

- Real-world applications of semantic graph analysis are used to discover relationships in large data sets.
- Graphs that represent interaction in semantic networks can become extremely large. In practice, the current size limit for graph analysis is  $10^8$  nodes, while the projected need is  $10^{12}$
- Betweenness centrality (BC) provides a way to identify critical components in a graph.
- The higher the BC value, the more "connected" it is to all other vertices in the graph.







## Data Intensive Architectures outperform typical supercomputers

	IBM BG/L 2005	Active Disks (2006)
Graph Edges	30 Billion	300 Billion
Graph Description	Random (Academic Problem)	Scale-Free ("real world" problem)
Processors and FPGAs	65,538 PowerPC 440	648 PowerPC 440 648 Xilinx Spartan 3 FPGA
Average Search Time	1.4 sec	218 sec
Level of Effort	6 months, 6 people on project, 2000 lines of C code	2 weeks, 1 person, 100 lines of SQL code

### Bi-directional Breath-First Graph Search Algorithm

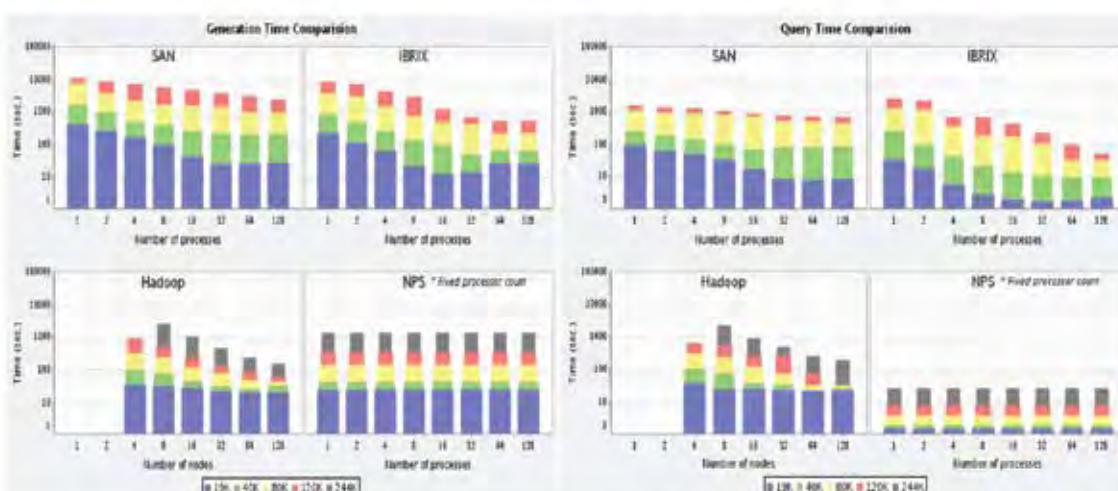
- 10X More Edges
- More than 12X Productivity Improvement
- 300 Billion Edge Problem Not Achievable on BG/L



Andy Yoo, et al.



## Excellent Performance on Related Problems



### Microarray Correlation and Analytics

- Generated 18TB for 1Mx1M correlation

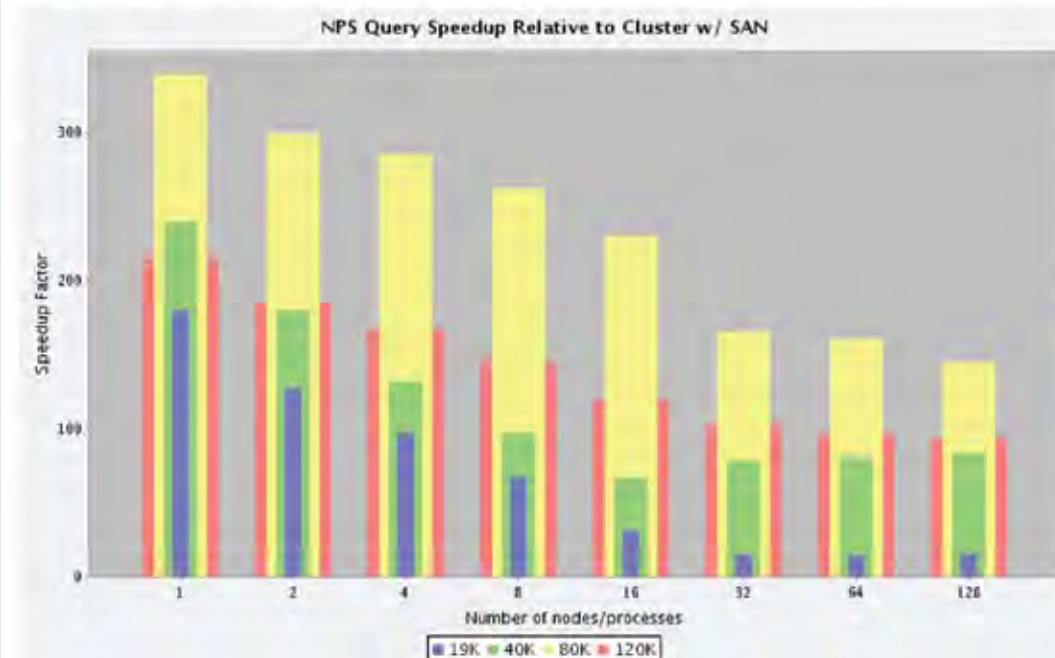
Delmerico, Chaudhary, et al., 2009





## Performance Relative to Cluster

- Hadoop and IBRIX speedups computed for equal number of nodes, NPS speedup compares entire machine to P nodes.



## Conclusion

- Large Amount of Data generated for semantic web and applications utilizing such concepts
- Many such problems are of national importance
- Traditional cluster and supercomputer architectures are not competitive
- CRL has deep expertise in data intensive supercomputing