14. Localization for e-governance

Prof Jitendra Shah,C-DAC Mumbai and Swapnil Hajare, Indictrans
E-mail : jitendras@ncst.ernet.in, indictrans@fastmail.fm

May I paraphrase a great thinker, *"Philosophers have interpreted the digital divide and language barriers variously. The point however is to break them."*

Kingmaker of IT Revolution: Reproducible, Widely Deployable Indian Language Content

Creating documents in local languages with ease is now well established. Use of free software for this is still only rare in e-governance . However, localisation in e-governance requires much more than that. It requires fluency in use of local languages in mass usage. Mass usage implies users from desktop to web based application, from individual and closed group to mass of users, from select machines to across the board variety of OSes, from unicode compliant systems to non-unicode, from affording users to non-affording mass users. For this to happen, an insider view of government functioning and its needs of software are essential. The localisation components will have to snugly fit into the government system, in particular into the currently used database and public interface systems. Going by the experience of the undersigned localised free/opensource has tremendous advantage in that there is the necessary freedom to adapt without costing a bomb. The national asset of software talent is worth little if it cannot or doesnot take on this challange. In the era of rising connectivity, content will be the king. Infinitely reproducible systems and applications in Indian language content may well be the *kingmaker*. If the national policy in IT in e-governance is aware of this need and of the specificity of this juncture, we are poised to leapfrog. If not, we will miss the bus. Internationally, people with various agendas, are watching us keenly.

Background: over ripe

What has happened over the last few years was the laying the ground for breaking the language barrier in the digital world. National and International state initiatives, voluntary movements, academic research, individual zeal and commercial interests have all propelled this change. This includes not only enablement and, proof of concept but also demonstrations and ready-to-use products. If one is not enamored of the latest fancy features, and can make do with a few limitations, there are many applications that can be straight away useful and can satisfy much of the needs of a desktop, database , www or a network.

Skew debate

As is natural , the initial attempts (say before 15 years) at bringing the indian language to computers resulted in some non-standard features. While attempts at national level standardisation were on, market forces played a major role. While the indian language adoption on computers was rapid , the technology was embraced as revolutionary. While input and printing in indian languages in limited environments grew , the computing in indian languages remained restricted. Over time, the technology based on ASCII came up against inherent limitations . This was reflected in abundance of content locked in as non-standard font-encodings. The content was anything but fluent. While the market for fonts (non-standard encodings) and converters (from one non-standard-to-another-non-standrad-encoding) flourished, communication, web-enablement,sorting,searching and globalisation of documents suffered. These contradictions were being played out in the market and on policy levels for last few years, mostly on proprietary platforms. Passions ran high even in academic debates. Often the debate was reduced to national pride versus multi-national conspiracy theories. This resulted in cacophony, more heat than light. The element of truth on both sides was getting clouded due to immaturity of technology.

GNU/Linux and Unicode standard

In the meanwhile, the Free/GNU platform was making strides on software front in many ways but a little slowly on indian languages. This slowness was steadily towards universal standards and hence almost deliberate.

Thus GNU/Linux community adopted unicode standard after a deep thought and could do so easily.

A wise decision : ISCII standard in the past, Unicode today

One must realise that about 10-15 years ago, when indian language technology was in infancy, and even internet and hence global communication was in

infancy, the iscii route was the only reasonable alternative as a standard. It has been used by many government offices. This has made a lot of legacy data easy to access today, in iscii as it is or by a quick (by almost trivial effort of) conversion to unicode. Now that globalisation demands fluent communication across many boundaries, universal standard is feasible and hence necessary. Transition from one paradigm to another is bound to be painful for some time. There are no two opinions about the necessity of a standard . Everyone agrees there must be one. And , currently there is no better than unicode.

Office suites in mass usage and government usage

At the user level perception, particularly at mass level, the most important computer application is the word processor and associated office suite . For a long time (till year 2002-2003), there was really hardly any alternative to the sleek and mighty proprietary software i.e. Microsoft Office ®. This is true in spite of many critics, in spite of many bugs and for that matter, in spite of so much cost. The emergence of OpenOffice.org was a real boost in the prospects of looking at GNU/Linux as a viable alternative OS on desktops of offices. The governments began to see GNU/Linux as a real alternative on many of its computers. There are some limitations even now. But surely, it is expected that these will be overcome . In particular, for indian languages , the strides made by Free/Opensource are significant.. In terms of total cost of ownership, the Free GNU/Linux based systems appear to be far better than proprietory.

Thus the current scene is overdetermined for an imminent sudden upsurge of use of Free/Opensource in indian context for IT for development.

The technological springboard

Over last few years, efforts like those of corporations (most notably IBM) have worked towards making many localisations technically and commercially feasible. These include establishing locale specifications i.e culturally specific invariants for a given language/nation/script , harping on unicode standard, supporting GNU/Linux vendors and supporting GNU/Linux as platform for many of its applications, ICU, ECLIPSE etc. For many young minds seeking career in software, 'Linux' started appearing as an alternative after all. Sun Microsystems support of OpenOffice.org made office users see a strong and supported alternative. Similarly support by many other international corporations have made 'Linux' a respectable name in corporate and government parlance. This appears to be the spring-board for a quantum jump.

As a passing observation, one is compelled to remind that in all this emphasis on 'Linux' and 'Opensource', one misses the 'free' (as in freedom) element of the GNU/Linux movement. But that remains the challenge for ethically inspired proponents of free software .

Free software in localization in e-governance

Localization in general, and that in indian language in particular, poses the issue of free software frontally. If the mass of indian population has to access benefits of IT revolution, the proprietor tags and associated price tags will prove to be the insurmountable obstacles. Freedom will be necessary condition to reach out to people. The sheer variety of cultures and variations in languages require not only changes in content (some caharacters, some terminologies, some writing systems, dictioneries etc) but also the tools and the basic systems. This is seen in areas like font creation, implementation of collation order, browser enablement in indian languages, cross platform compatibilities, easy reproducibility, tweaking of the system and most importantly informed local support for it etc Economic models for viable local support will have to evolve. The government policy will play the major role in this.

Indictrans Effort : mass deployment and government site opportunities

Till recently, the undersigned was convening the localization effort through a non-profit, un-organized group of trainee-volunteers called Indictrans Team. What indictrans team has done over the last one year is the emphasis on deploying the free software in e-governance and to a slight extent in education and rural context. Principally our aim was to identify stumbling blocks in adoption of the free localized opensource software in e-governance. More importantly we looked at the difficulties in government-citizen interface. Team looked at the kinks

in adopting unicode standard in applications developed for government functioning etc.

This may be seen in contrast to many others who started much earlier than us and have been doing the yeoman's service to indic localization. They are completing the important tasks of localizing GUI's like GNOME/KDE or some applications.. See www.indlinux.org site for more details of language teams. While users who would want a complete GUI in local language are growing, they will significantly impact the use of software only after a few years. There are teams working on some even more ambitious projects like making machine translations from one indian language to another etc.. The machine translation, we expect will mature only after a few years , i.e. after massive corpa are analyzed and lexicons built. These tasks require much greater resources, skills and much deeper commitment. Many enthusiasts have been working on these projects with frugal resources. Indeed, Indictrans team always acknowledged that they stood on the shoulders of these teams.

Following describes briefly two major tasks accomplished by Indictrans team in the area of localization. Following also enumerates various IT components developed by Indictrans in the process.

First was standardization i.e. conversion of live data and file-journey-management database from non-unicode to unicode standard . The second was Voterlist search engine for Chief Electoral Officer of Maharashtra. The components include , fonts, writing systems, filters/converters with sample conversions for important national IT projects, sample localised applications, some content, bootable CD and a building a motivated team etc.

I Standardization : Conversion from legacy silos to unicode

when we started in August 2003, we decided to chose govt. Offices as the starting point for implementation of our technology as we saw a lot of potential of interacting with community through this channel. The biggest obstacle that we could see was the huge silos of data already created in older technology. This was preventing some beaurocrats who were positively inclined towards use of Free/Opensource solutions in their offices like direcorate of IT Maharashtra, in dopting the solutions.

We therefore decided to first provide solution for converting this legacy data into new open standard i.e. Unicode. The Deptt. Of IT Maharashtra already had a system in place called DJMS (Document Journey Management System). It was a browser based system supported by documents and a database of metadata for these documents.This application was used for transparently tracking each and every document as it passes from one table to another. Their major problem was making this available in local language i.e. Marathi in an affordable manner. They also wanted a standard solution to manage Marathi content atadministrative level (i.e. Database level).They were using ISM/ ISFOC as solution for Marathi. Unicode sounded perfect solution for them as it offered minimal reengineering at administration level and affordable alternative at user level.

We undertook the job converting the documents from ISFOC to Unicode using a converter developed by us. The files were of various formats like doc,xls,rtf, html. There were about 6000+ files (see table showing distribution of file types on indictrans in Conversion page of www.indictrans.org). We also converted the database which was based on DB2 so that the documents could be searched in Marathi using the existing application.

As a result of this conversion, DIT can now use opensource applications like Openoffice.org and Mozilla across different platforms. All this was completed in less than 6 months as a result of which a letter of appreciation was conferred us by Directorate of IT, Maharashtra.

The experience is significant in that it shows that massive silos of legacy font-encoded content can be converted to the standard (unicode) and that this can be done without much disruption of running system. All it needs is the will (of the authorities) and perseverance (of the project implementers).

In the above snapshot one sees a page with many ilegal characters in original document.

II. Voterlist Search implementation

In May 2004, while, we had several technologies and tools, we were looking for some challenge where these tools could be demonstrated in mass-deployment . Just then we saw a huge hue and cry in the press and electronic 'media about missing names in voterlist during the Lok Sabha elections concluded in April. We announced our intention to use unicode standard to overcome the problem (see Indian Express marathi publication dated 9 May 2004). Maharashtra state assembly elections were due in a few months (October 2004) .

We , proactively, and without much knowledge of the state of art in electoral roll computerisation, approached the State election authorities. Naturally, the election authorities were skeptical, but open and positively inclined to explore. However, the undersigned was also told by veterans that many attempts at the voterlist searchability had been made and failed and we may be yet another attempt in the same line..

We knew that we had a different approach than most others and hence we dared. We had the technology to break the language barriers in the long chain of the process of voterlist making and could make the access to the list truly universal. In voterlist search engine implementation, in effect, we broke the language barriers in a mass deployment application.

The original data was in iscii standard. We made the data available in unicode standard, searchable using modern tools and rdbms (pgsql, tsearch2). We made display available in unicode as also in non-unicode. Thus we allowed access to latest as well as oldest OS. We provided interaction across the net . All this while fully working with free/opensource. We must admit that we used 'dynamic font' technology as a last resort to support computers where the user may not even wish to install the free ascii based devanagari fonts on MSWindows legacy systems. .

We demonstrated that whether on desktop or on the web, whether on new (Unicode compliant) or old, whether a rich or a poor PC holder, (refer to licensing costs),

there is no barrier to work in indian language. We provided a transcription (transliteration using itrans) from english to indic language and vice versa .We have also demonstarted the interoperability i.e unicode and non-unicode (whether ISCII standard or non-standard) content can be seen by each other and interaction is possible across the divide.

While our software was put up on the web as free software under GPL, the implementation was done for the Chief Election Officer of Maharashtra by using an agency already in rate-contract with the Maharashtra Government.

The main features of software used for this search engine include:

- Full Text Search can be made by name and/or address OR Photo Identity Card No.
- Search strings can be a number of strings , and search is irrespective of order.
- Most important feature is that spelling errors are broadly tolerated which helps search .Take the example of chellamma. This is spelt in Maraathi in several (12 spellings , see snapshot) different ways. This is also dependent on the variations introduced by a marathi speaking person hearing and then writing a south-indian name in devanagari. and typical names like Krishnamoorthy (12 spellings) and Aggarwaal (16 spellings) or gayakvaad (24 spellings).
- It is possible to search giving partial strings of the names (with a wildcard character as suffix)
- Search results are ordered as per collation order (first collation order of first string, then of second string etc.).
- Search results include Polling booth name,address and Other info such as serial No. and Part No of voterlist.
- completely based on Free GNU/Linux (Knoppix3.4+Apache2+PHP4+Postgre SQL7.4.3)

- original ISCII data in .DBF format was converted to unicode and imported into PostgreSQL
- complete source code of the search engine is released under GNU GPL on savannah.nongnu. org as also on ww.indictrans.org (see vl2-libs)
- copyrights belong to Swapnil Hajare & Sreepathi Pai but license is GPL . Thus any one can use the software without any restriction.
- powered by intuitive similar sounding search technique called "souindics" developed @ Indictrans
- Although the data itself is in Marathi search can be made by input in Marathi as well as English
- Search results can be in Unicode Marathi/ modified ITRANS English OR in legacy fonts like Shusha,DVBWTTYogesh.
- Results are also available in Dynamic font AkrutiHnbDynamicYogini (courtsey: Cyberscape Multimedia Ltd.)
- Search time less than 15sec on entire data on an average
- Search results can also be obtained in the form of voterslip as PDF with embedded Shusha font which can be viewed in any PDF viewer without installing font (note it is not possible to embed opentype font in .pdf as in October 2004)
- Though the implementation was for Marathi, there is hardly anything in the code that world need to be changed for most other indian languages, except the converters.

The software has been accepted by CEO(Maharashtra) as part of Election Commission, after serious testing by C-DAC Pune. The coverage of pilot at Mumabi+Thane was for about a 1.25 crore (12.5 million) voters. More than 7 lakh hits were recorded in 10 days till election on 13 October. This is apart from helplines in Collector's offices also using the same software on desktops. For more details see voter list page on www.indictrans.org .

Future applications and improvements

Future applications include usage of voterlist at district levels for municipal elections or local government elections. The same software can inform the registration process to eliminate duplicates. The utility owners like phone service providers or electricity providers or mobile phone providers can remind their subscribers with message for voting, polio pulse or some such message with some location specific info like booth details. In case of accidents police departments can use the seach engine to locate addresses or family members. Combined with name transliteration program, the search engine can be used for multi-lingual search. Connected with a GIS database, the same can be used for delineating municipal wardboundaries, booth-location planning, other regional data etc. Most of these applications require large multiples of deployment and indian language abilities are essential for the applications. The databases and GIS etc all have to be compatible. These can be built easily on free software platforms.

Going beyond voterlist, same technologies can be used for land records, telephone utilities, electricity utilities, insurance utilities and other mass lists.

The search engine can be imprved to work with larger data with greater efficiency. The phonetic search can be improved with more knowledge of occurrence frequency of names etc. The search intelligence could be improved using region specific sounds e.g. Yamuna/ jamuna. Vipin/bipin etc.

III. Components of Localisation: Felt needs

Apart from the voterlist search program ,Indictrans team came out with the following broad range of tools/solutions towards localisation as response to requirements of users who wanted to use localised software.They are also team's perception of the felt needs, among others, in the direction of mass deployment of localised software. These are, therefore,also candidates for componentisation for pluggable, reusable modules to be part of mainstream.

Since Indictrans is not a commercial entity, it doesnot claim have a sustainable economic model of activity nor are the follwing projects finished or packaged . Indictrans aimed at only proving the concept and learning the nitty-gritty. Indictrans also didnot carry out exhaustive surveys of current commercially available solutions or freely available techniques .

1. System level Localisation : Indictrans helped in Locale specification and Translations of

GNOME messages in Marathi and Gujaratii. Indlinux have now a more comprehensive repository of indic locales.

2. Font: Indictrans created and maintained fonts for Devanagari and Gujarati.These were Unicode opentype, now available on TDIL website. Indictrans displayed performance of all these fonts, as snapshots, on all platforms on www.indictrans.org.

3. Conversion: Silos of content generated in font-encoded form remains a problem with many institutions. Conversions of these content from legacy font-encoding to unicode is essential to make way for opensource OSes to be used. The main technical difficulty is that such content is in many file formats and is embedded in live databases. Then, owners of data may have concerns of security of data as it may be sensitive data. It requires a visionary leadership at user end and confidence in technology on the side of service provider. Indictrans tried to tackle these problems as follows.

   a) Undertook a major job for Government of Maharashtra (see letter of appreciation www.indictrans.org http://www.indictrans.org/src/letter_of_appreciation.jpg also described above)

   b) Recently converted HTML files of a culture and art oriented website with thousands of webpages. to Unicode (This was as per suggestion at Pune Localisation Review meet around August 2004, organised by TDIL,)

   c) Conversion of sample of land record data of Bhoomi project for NIC Karnataka,

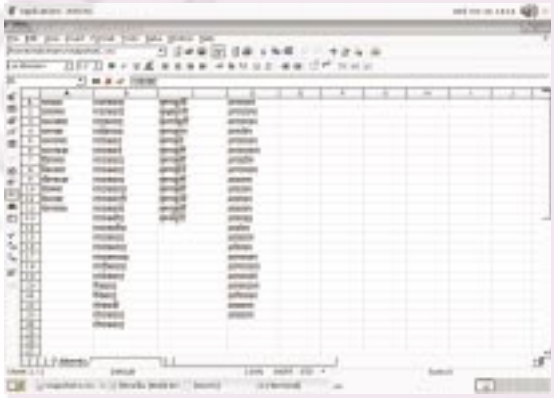
   d) Similarly Smple data for rajasthan Raj Corporation

   e) Conversion of MarathiWorld.com site (1800 pages from Mithi font to unicode)

   f) Conversions of sample files from Akruti font to uniocde for IITBombay

4. Inputbhaaratii: Several applications for inputting indic scripts on web were developed. Software implementation of Inscript on Java (without the need for a plug-in like ISM), as also IIT Mumbai's KeyLekh layout (see http://www.indictrans.org/typebhaaratii/) Thus anyone can use web based indian language without the server having a commercial plug in. And the text so inputted is saved in indian languages in unicode.

5. naamabharatii (name transliteration see http://www.indictrans.org/naamabhaaratii/) : Many offices /lists have indian names in usual/conventional english spellings. These need to be converted to indian language proper spellings before they can adopt indian language systems for say database or for correspondence. It was found that there was no solution on opensource. The need was encountered in Govt of Maharashtra as also LIC etc. A program using dictionary+hueristic was developed for several languages and is available on indictrans site for demo.



Actual variations of spellings in some available records.

6. Application Localisation: Horde, a messaging framework has been translated into marathi and localised.

7. Developemnt of bootable CD for localised GNU/Linux: gnubhaaratii see tutorials on the site in Hindi,Gujarati, Marathi and English,

8. Localised GIS : grass, a free GIS map being used to put time series data of state election commission on panchayat wise map of Maharashtra. Only a beginning has been

made.This involves using pgsql database linked to mapping grass program. Clickable, zoomable maps for any geo-referenced or geographical database would be created.

9. Localised software for Rural Development: The undersigned was a consultant to Tata Institute Of Social Science,Mumbai for their Rural Campus (500 Km south of Mumbai, called Tuljapur). A series of localised applications were developed for induction into the development and curricular (B.A. in Social Work) work of TISS. The setting up of on site network, establishment of localised OS gnubhaaratii with localised applications (like Openoffice and Mozilla) and training the faculty and other staff was undertaken and completed along with training. Power availability at the make-shift campus remained a daunting problem. With this experience, while talking of IT in rural areas, Indictrans team cannot forget the importance of stable electricty.
10. Localised geometry: DrGeo is a program on interactive geometry. It has been localised with tutorials in Hindi, Marathi, Gujarati,
11. CBSE IT Curriculum: Indictrans team attempted to enable implementation of CBSE curriculum in indian language using openoffice and other free/opensource software on GNU/ Linux. The implementations on OpenOffice are of corse cross-platform.
12. Localised Editor: Yudit has been localised in Hindi, Marathi and Gujarati
13. Team building: Last but not the least, Indictrans shows an example of motivating and cultivating a team of computer science graduates along with a bunch of other graduates to work on localization. To sustain such a team with frugal means, while the IT job market is booming, has not been easy.

Change in quantity leads to change in quality

With the introduction of GNU/Linux and Open standards like Unicode in e-governance, the nature of e-governance applications have changed a lot.It is now more open for participation by users and more obvious to operate as against an imposed responsibility from management. This has opened the possibility of participation by much larger numbers in governance, whether at government or at private corporate levels. This openness also fits nicely into the requirements to fulfil the guidelines mentioned in recently introduced 'right-to-information' act . This increase of numbers will surely affect the very nature of communication and may lead to qualitative change in sharing of power among stakeholders .

Future trends

The localised e-governance will soon penetrate deeper. When all base information (from district gazeteers to routine statistics) and news and other info becomes available in local languiages on standard, open platforms, the demand for information will spread. Much of the information will be geo-referenced. Linking GIS to alpha-numeric data will be crucial. The wireless technology is poised to leapfrog. Hence Indian language content, connected with wireless and mobile technology will make participation of large mass in governance far more informed and hence effective. Language technologies will play a major role in these trends. More so if they are free as in freedom. Even more so if they are reprducible at will i.e without any licensing issues.

Conclusion

Localised e-governance is not just a technical freak but right of the people. It is now feasible and hence a must. As for the alternatives, in particular the alternative of proprietory loaclised software, freedom as well as cost factors will favor free/opensource. When considering only cost, it is essential to look at total cost of ownership for the nation as a whole and see beyond the mindset barriers. With that perspective free/localised software is the obvious, feasible and perhaps the only choice for mass of people to be benefitted from the IT revolution. If not the digital divide will yawn further.

Maharshtra DIT : Conversion to Unicode Project: A sample of page in original fontencoded format