Marathi and Computer

Alka Irani
Chief Investigator IndiXII, Senior Research Scientist, C-DAC, Mumbai, alka@cdacmumbai.in
with implementation support from Swapnil Hajare et al of
janabhaaratii (funded by TDIL, MC&T, Govt of India)
C-DAC, Mumbai (formerly NCST) • janabhaaratii@cdacmumbai.in

UNDP defines human development as expanding the choices for all people in society. This ensures the creation of an enabling environment in which all can enjoy long, healthy and creative lives. IT can be used to promote education, learning, communication and networking, neither technology nor products can be taken off-the-shelf from other countries, because the conditions, needs and languages are radically different.

Given the present financial, technical and linguistic constraints there is an urgent need to improve the conditions for equitable and affordable access to computers so that the benefits of IT are distributed across all sections of the society.

It is said that 'No man is an island'. In the same way, according to me, the days of stand-alone computers are over. Computers all over the world are networked to form Internet. This new, dynamic, flexible, powerful medium called `Internet' is dominated by one language. According to a survey, 85 % of the contents on Internet are in English while less than 5 % of people speak English as their first language.

Collaborative content development for Marathi – The need

We Indians are proud of Eastern Civilization but are in great danger of losing the advantage our rich civilization has if we don't preserve our cultural heritage.

Language and culture go hand in hand. Currently, Indian software professionals are busy making unbelievable progress on all fronts of computerisation all over the world but as far as our own cultural repository is concerned our contribution is naganya (insignificant).

Marathi, computer and Internet

Marathi is a language with rich heritage and abundant literature. For most of the people in rural Maharashtra (educated or uneducated) this is the only language of communication. Therefore when one thinks of using computers for public health, rural development, education, e-governance and media along with enabling computers to read, write, print Marathi text, the interactions with computer, the help material etc. and most important the contents to share among community members need to be in Marathi and in the script used by Marathi i.e. Devanagari.

We can see quite a few websites in Marathi. Most of the Marathi newspapers – Loksatta, Maharashtra Times, have their e-copy on Internet.

A question to ponder over is how many Marathi speaking people are using Computers for interaction (chat) in Marathi or for sending mails (e-mail) in Marathi.

Education and Computer

The conventional education system is unable to cope with the rapidly changing information needs of the education domain. The vast Knowledge Base available in foreign languages can reach masses only if it is made available in local languages. Currently, the educators are unequipped to handle the rapid pace at which the Information world is moving. If the current model of teaching is to be sustained continuous input to the educator's knowledge and a continuous collaborative learning mechanism must be available to them. Getting familiar with this new medium will make their job easy and enjoyable. Once they have their lessons neatly organised and kept, it will not be necessary to go through the process of preparing presentation all over again. They can make incremental changes. Many of the mundane day-to-day jobs like keeping track of students, their activities, progress can be minimal. Email is a cheap and affordable medium for information exchange while web-pages can be used for publishing as well as for bulletin boards.

Computer, local languages and Employment opportunities:

Ample employment opportunities are available, the need is to make local languages like Marathi successfully to computers. Many Government offices are confronted with the task of working in Indian languages. The nationwide project PURA on initiative of our President and a great visionary Abdul Kalam for Providing Urban amenities for Rural Areas is already working in that direction. If competent people (teachers/students) are gathered, and resources are provided, educational contents for rural settings can be developed using local skills. Contents will be understood much better as they are written by the people having similar cultural background (peer learning). Local newspapers, magazines/TV channels need lot of contents to be input, edited and produced in local languages.

Computer and contents

Computer is an ideal medium for content management. With drastic reduction in costs of disk drives and invention of many affordable, flexible, compact storing devices like pen drives, the contents can be stored efficiently in compact format in a way they can be accessed anytime, anywhere.

Some desirable characteristics of contents are:

- Contents should be in standardized formats. For a collaborative framework to succeed, the methods and conventions used must be stable, universal and scalable.
- Multiple ways of organising, viewing, browsing data must be possible
- Multiple layers (that can be merged or separated as per the need) building various hierarchies, abstractions must be possible.
- Contents should be attractive to view (education should be fun). With use of multimedia it is possible.
- Contents management systems should have provision for entering/retrieving contents for disable people. This means visual or speech interfaces should be possible.
- Content units should be distributed at the same time fairly independent to allow content developers work independent of Computer Network when it is not accessible

The Scenario

Many stand-alone packages for bi-lingual text input exist. The two important landmark achievements in 80's from CDAC are:

- Rupantar (Software solution)

The rupantar phonetic coding scheme was designed during early 1980s at a time when there was no Devanagari text-processor available for inputting the text; the PC revolution had not caught up and most of the work was done on main machines of Digital at NCST.

It was designed during development of transliteration softwares Swaroop and then Rupantar which were basically computer-based systems for converting names written in English script into Devanagari equivalents.

Transliteration of the entire English Telephone Dictionary in Marathi, sorting and made it ready for printing; also used the scheme for printing cheques, printing certificates, etc.

- GIST (Hardware Solution)

The Graphic and Intelligence based Script Technology (GIST) group of C-DAC, Pune has introduced a GIST card for inputting Indian characters on DOS platform.

The GIST card has been the cornerstone of some of the most crucial computerization programs of India such as the Land Records Program, the Election Commission Identity cards, and citizen surveys.

- Standardisation

Fundamentally, computers just deal with numbers. They store letters and other characters by assigning a number for each one called encoding of the character. Currently there are two standards for devanagari encoding on computers.

ISCII – ISCII is an Indian national standard for character encoding which was revised many times. The ISCII-91 code retains the standard ASCII code while utilizing the upper ASCII codes for Indian scripts. This makes it feasible to use Indian scripts along with English computers and software in an 8-bit environment.

The ISCII code table is a superset of all the characters required in the 10 Brahmi-based Indian scripts including Devanagari. These scripts share a large number of structural features between them as a consequence of their common Brahmi origin.

- Unicode: character encoding

Before Unicode was invented, there were hundreds of different encoding systems for assigning these numbers. No single encoding could contain enough characters. These encoding systems also conflicted with one another.

Unicode provides a unique number for every character, no matter what the platform, no matter what the program, no matter what the language. The Unicode Standard has been adopted by Apple, HP, IBM, Microsoft, Oracle, SAP, Sun, Sybase, Unisys etc. Unicode is required by modern standards such as XML and Java. It is supported in many operating systems, all modern browsers, and many other products. The emergence of the Unicode Standard, and the availability of tools supporting it, are among the most significant recent global software technology trends.

Incorporating Unicode into client-server or multi-tiered applications and websites offers significant cost savings over the use of legacy character sets. Unicode enables a single software product or a single website to be targeted across multiple platforms, languages and countries without re-engineering. It allows data to be transported through many different systems without corruption.

(Unicode and ISCII standards differ. ISCII is a 8-bit character encoding, Unicode 32-bit character encoding. Unicode was based on ISCII-88.) The latest ISCII Standard is ISCII-91.

However, even today many standalone packages – text processors for inputting text and printing with their own encodings.

What was done to enhance quality, inclusiveness, diversity at NCST (now C-DAC Mumbai)

The two recent projects at NCST (now C-DAC, Mumbai) is a step in right direction to make Indian Language pro available to masses on computers and Internet.

1. IndiX enabling the medium for reading/writing Marathi

The basic IndiX agenda is that the text processing should be as easy as it is for English user. The technological challenges addressed by IndiX is to identify the minimal, logical and required changes in Indic text processing and embed these changes within the lower most level of a widely used and deployed software architecture GNU/LINUX.

GNU/Linux (many times known as Linux) is a computer operating system and its kernel. It is one of the most prominent examples of free software and of open source development: unlike proprietary operating systems such as Windows and MAC OS, all of its underlying source code is available to the public and anyone can freely use, modify, and redistribute it.

A Linux distribution for example Debian Knoppix bundles large quantities of application software with the core system, and provides more user-friendly installation and upgrades.

Initially, Linux was primarily developed and used by individual enthusiasts. Since then, Linux has gained the support of major corporations such as IBM, SUN, HP and Novell for use in servers and is gaining popularity.

2. Localising the medium for Marathi - janabhaaratii

The broad objective of janabhaaratii project is to enable wide use of Indian language computing through Free/Open Source systems and applications localized in Indian languages.

Janabhaaratii teams responsibility is to promote use of Indian language computing among people through development, deployment and support to the community. All this using free and opensource softwares which can be used, modified and distributed free of cost. As part of the project, the team has come up with a LiveCD (a cd containing complete GNU/Linux Operating System) which is out-of-box working solution for Indian language computing. Using janabhaaratii liveCD, one can have all the tools required for Indian language content creation available at no cost. The CD is also supported by good HowTos and guides.

Another initiative by the team is to make more and more people aware of the Indian language computing through the use of freely available tools with the help of awareness workshops and training programs. The team has organised/participated in a number of workshops and trainings for educational institutes as well as govt departments.

As part of community building, the team has also developed a collaborative portal called aantarabhaaratii as platform for all the content developers, translators, engineers, developers etc to come and interact. Marathi localisation of popular desktop applications is also being coordinated with help from marathiopensource group. With more and more volunteers contributing in translations, we aim at developing marathi interfaces for most commonly used applications like office suite, web browser etc.

Here we quote Sowa

The point is not just that we can handle large *chunks of knowledge as though they were atoms, the important thing is that we should be able to find out way through these complex nested structures to whatever individual fact or relationship we might need at any given time. We should be doing this in a very flexible and efficient way and avoid having to look individually at each of the vast number of facts which are not relevant to the problem at hand.*

— Sowa[Sowa, 1984]

Framework for Content Management: A necessity

The framework of information processing by modern computers are still not so flexible compared to human flexibility in information processing in the real world where many problems are ill-defined and hard to describe using algorithms. Therefore in order to cope with such real world problems it is essential to pursue the fundamental ways of human-like flexible information processing. The buzzword here is “Real World Computing”.

“Real World Computing” is a paradigm of information processing which aims at furnishing the real-worldness (or flexibility) of human information processing to information systems.

Human beings are “Real world

Information systems"

In order to make "Real World Computing" possible following issues need to be tackled.

- ❑ Cost of service
- ❑ Ease of access to information
- ❑ Reusability of information
- ❑ Empowerment of the typically bypassed to use ICT
- ❑ management of information overloading (too many contents)
- ❑ What needs to be done?
- ❑ Provide access points (community centres/kiosks) (infrastructure)
- ❑ Allow mechanism to transfer knowledge
- ❑ Promote easy, affordable and enriched medium
- ❑ Create frameworks to share contents in local languages

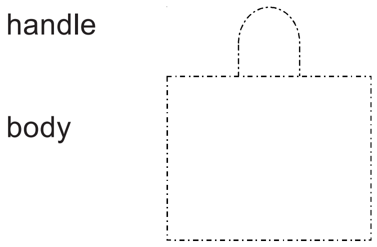
Hardware costs are falling day by day. With the advent of OpenSource and availability of GNU/LINUX platform the basic software has become affordable. With Government of India and various state Government agencies sponsoring various projects and building infrastructures, providing access points or community kiosks, success of collaborative framework depends upon creating frameworks to read/write and share contents in local languages. The first step towards creating the framework is enriching the medium called computer.

Enriching the medium

In order to be effective, the Knowledge Representation Framework on computer medium should provide ways and means to the one who designs the knowledge base to state what it is all about in an unambiguous way to enable the system to search relevant information. To state this it needs vocabulary. This vocabulary cannot and need not be created arbitrarily. In order to have a universal convention, the concepts from natural languages should be the building blocks for computer-based systems as well.

The proposal: Intention-based knowledge organisation

Basic units for information exchange we are proposing are chunks.



Chunk consists of two parts handle and body.

A handle helps in retrieval of the body of the chunk. Handles contain info and meta-knowledge or what is called 'intention' of the data. The body is called 'extension' of data.

In neuroscience, the basic unit of human memory is described as a chunk. (note: While proposing production systems as general models of cognitive architecture {Anderson, 1983], [Newell 1990], and as a way to represent human expertise in computer programs, the primitive structure assumed is also chunk.)

Retrieval in computer systems:

If we want the system to be flexible, retrieval should be possible in many ways: deliberate, through meta-knowledge, multiple scanning and searching. Picking data using 'handles' or organizing data based on 'handles' can make retrieval fast and effective if 'handles' are chosen carefully.

Chunks can be organised using abstractions

- leveling
- partitioning
- pointing
- clustering

Implementation of chunks is simple just create text files with tags. (XML format)

For example: < Name >

< author >

< creation date >

< body >

Chunks can be classified. Each class can be described using what is called dtds.

General purpose retrieval

The biggest advantage of computer is its ability to transform/rearrange/collate/ reorganise data on demand to suit a purpose. One need not 'hardwire' the data right in the beginning. Once the 'handles' are identified, the same chunks can be organised in various ways depending upon the purpose.

For example: Marks can be arranged studentwise or subjectwise.

Books can be organized authorwise or subjectwise or by year of publications

Conclusion

Various groups are advocating different formalisms as "best" for the job. Many times issues are handled in isolation. Methodology and tools are not provided in all the cases. We have no perfect answer to what is the best structure. Things will evolve. The structures will emerge.

The tasks of building the medium and developing contents can be taken up irrespective of the final structure if we can start building chunks.

An enormous amount of work is required to build an underlying vocabulary.

However, various sources, various Government-sponsored projects like coil-net have already started and we all can join hands and work.

Action plan :

1. build dictionaries – work on standardisation
2. build dtds – data definition structures for various tasks in various domains
3. build linguistic resources – concepts – higher level concepts, - vocabularies in specialized domains
4. build contents as chunks (meta-knowledge + knowledge) without finalising structuring, presentation choices

If all the "content developers" i.e. teachers, trainers, library science professionals, writers, reporters contribute to the "enrichment of Marathi on computers" by filing the chunks, Marathi and Computer will certainly go hand-in-hand.

Immediate priorities include

- More and more content development in Marathi using a content management framework
- Translations for application GUI and operating system messages

Acknowledgement :

This work was part of my Ph.D.(BITS Pilani) carried out at NCST (now C-DAC India). I am thankful to my colleagues both from Indix2 team as well as janabhaaratii team for enabling the platform for local languages. I am thankful to Gracy Abreo for proof-reading this text.

References :

1. 'A Unified Model for Concept Structuring' – Alka S Irani, Ph.D. Thesis , BITS Pilani 1995
2. 'Texts across contexts: Indix encoding and its applications', Sylvia Candelana de Ram(Comp Research Lab, New Maxico State Univ.), Alka Irani and M Srikant in ACH at Fordham Univ in Bronx. NYC, June 1990
3. 'Conceptual structures: Information processing in mind and machine, Sowa J F, Addison Wesley.

☆☆☆