NEED FOR MULTILINGUAL DATABASE IN INDIAN LIBRARY AND INFORMATION SYSTEMS

Smt. V. Rama Choudary, Smt V J Suseela, Smt. V Uma

ABSTRACTS

Need for creating a multilingual database in an Indian environment is emphasized. Discussed various standards and efforts relevant to both multilingual processing of Non-Roman Script and processing of bibliographic databases. Suggested for multilingual software suitable for Library and Information System, highlighting the necessity of active involvement of NATIONAL LIBRARY and INFLIBNET for accessing information through Network.

Automation of Libraries in India, though commenced with Special or R&D Libraries and enjoying the fruits of it, Academic Libraries, at present are gaining importance with the support of INFLIBNET PROGRAMME OF Govt. of India with regard to Computerisation of their Library operations and the bibliographic databases. India, being the country with multilingual states, all the Educational Institutions, including Institutions of Higher Education emphasize on the study and research of Regional Languages. Hence, the literature and relevant Information should be available to one and all, in all the languages across the barrier. Thereby, the Computerisation of database for the benefit of user's information search would be complete and functional.

The following problems related to the automation of regional languages database are worth mentioning and due to which many of the Indian Libraries are yet away from Computerisation of these programmes:

1. Difficulty in feeding the records of regional languages in original script as the key board operation is in Roman Script.

2. Though some libraries opt for transliteration procedures into English as an alternative to enter their data pertaining to Indian/Regional languages, there are certain limitations with regard to standardization of spelling/ usage while inputting the data.

3. Ultimately the user encounter the same difficulty, while retrieving the Information whereby loses the precision in retrieval of the data.

4. The last but the most important is application of the multilingual text/script in the creation of bibliographic database and also using the same on Library Information System Software is still in the primary stage.

Undoubtedly, it is the Roman Script, i.e. English, has emerged as the defacto language for Computerization and also being used through various communication channels, but the various efforts of ANSI, ISO and ISI, in the line of standardization of the various language scripts/ character-sets suiting to the computer environment, are to be stated in this context.

ASCII CODE

The American standard code which is most widely used character-set for information interchange whose international equivalent is ISO 646. The standard that controls the mixing of character-sets in bibliographic records is ISO 2022, Information Processing. The disadvantage of this methodology for mixing scripts, is that to know which character is represented, and also to know which characterset is in effect.

ISO/IEC 10646

The Universal characterset (UCS) developed by ISO in 1993, is designed to accommodate all scripts in a single characterset. This is structured as eight groups of 16 planes, each plane containing 256 character codes. The basic multilingual plane (BMP) is the first plane of the characterset which include Latin, Greek, Cyrillic, Armenian, Hebrew, Arabic, Devanagari, Bengali, Gurumukhi, Gujarathi, Oriya, Tamil, Telugu, Kannada, Malayalam, Thai, Lio, Georgian, Japanese, Kana, Korean, Hangul, and unified set of East Asian Ideographs. It not only supports modern communication, but also classical forms of languages such as Greek, Hebrew, Latin, Pali, Sanskrit and literary Chinese, based on which Unicode Standard was developed by a Consortium of Information Processing and Computer Companies with the aim of encoding the characters used in written communication both modern and historic. The Unicode Consortium are building Unicode-based software, for Library applications.

NON-ROMAN SCRIPS IN BIBLIOGRAPHIC RECORDS:

These scripts can be incorporated into bibliographic records in various ways.

a) The single script record is simultaneously providing conventional Latin Script access and the other to the Non-Roman script.

eg: Dobis/Libis system of technilib in Melbourne, Australia which has records for Greek works both in Latin script alone (i.e. Romanized records) in the main file, and in Greek script (supplemented with Latin script) as a local file.

b) Multiscript records suggest the use of a single multi script characterset.

eg: ALEPH (Israeli system) which supports five scripts (Hebrew, Latin, Cyrillic, Arabic and Greek). Any two charactersets may be displayed together on any one display screen.

c) Multiscript records with characters invoked within the field contents. It has the greeted flexibility and the ability to designate and invoke a characterset within the contents of a field (in case of Non-Roman scripts).

This option is used in both. Language Experts persuaded the Library of Congress to provide original script cataloguing on cards, for material in these languages as they felt that the Romanization was inadequate, and the correct script could be formulated in machine-readable form.

A USMARC record that includes Non-Roman Data is structured as a completely romanized record, supplemented with data in the scripts of publication. USMARC and UNIMARC is also used by the British Library while processing their records in Non-Roman Scripts. Whereas inclusion of Non-Roman in USMARC records is optional.

In the process of creating a Non-Roman database, certain problems such as bidirectional scripts and sorting were posed and resolved as follows:

1. Bidirectional scripts

Arabic, Hebrew and Urdu are written from right to left when the default direction of any standard code i.e. US MARC, UNIMARC, etc. is from left to right. To resolve this problem a provision is made in the field as /r. Apart from this IFLA has developed additional guidelines for ISBD punctuation in Bidirectional Context.

2. Sorting the data

Sorting of multiscript data is still an unresolved issue. Wellisch is of the opinion that useful order may be obtained by a large family of script with further subdivisions based on the volume of book production and it is not also possible to device a single sequence for all the characters in all scrips that are universally accepted.

INDIAN SCENARIO

India is a country, with multilingual states each representing basically a State Official language in addition to other local languages in some states. On the whole India has approximately 250 languages and 800 dialects. Among them, India recognised 15 State Languages with Hindi as the National Language. However, Sanskrit is considered as the Classical Language. Under these circumstances, many Indian Software Systems are making efforts to design multilingual software hospitable to various Indian language Scrips. The important software, presently in usage is the GIST Technology and recently Central Institute of Indian Languages, Mysore has developed a Indian Script Utility and Script processor for Hindi, Kannada, Marathi, Sanskrit, Tamil and Telugu.

1. GIST (Graphics and Script Technology) chips is a high resolution monochrome, bitmap, graphic CRT Controller. It was developed by

CDAC (Centre for Development of Advanced Computing), Poona. It has evolved basically from one of its major applications namely Indian Script Processing. The chip supports interface to 68008 motorola and direct TTL monochrome monitor. CDAC recently released LEAP : a multilingual English-plus-all-Indian-Languages word processor with many advanced DTP features like type fonts and page make-up aids. GIST key board layout is presented in Annexure-2 (item No.1). It is mainly featured to Information Systems, data processing alongwith other applications related to communication.

2. Indian script utility is a systems software intended to serve as a tool for developing any kind of application software including educational ones through the medium of Indian Language on IBM compatable computers on MS-DOS operating systems. The BHASHA Software has been designed by Central Institute of Indian Languages, Mysore, in such a way by using this Indian Scrips can be displayed in the same way as Roman Script. This is available in BHASHA Diskette and the key board layout is as recommended by the DOE and it is also corresponding to ISCII in Indian Scripts. (Ref: Annexure 1). The software accommodates toggling facility from Roman Script mode to Indian Script mode and vice versa by operating <F9> key and default being the Devanagari (i.e. India, Sanskrit and Marathi). The following <alt> Keys move the script to different languages viz.

<alt> and <1> together Devanagiri
<alt> and <2> Kannada
<alt> and <3> Tamil
<alt> and <4> Telugu

Mainly it supports text processing in Indian Scripts alongwith transliteration facility from one Indian Script to another. Apart from these there are also different keyboard layouts for different languages available. A sample keyboard layout for Devanagari Script is given in Annexure 2 (Item No.2)

3. V M 4005 :

Multilingual Display Terminal package developed by VXL Instruments Ltd., Bangalore is a tool for using various languages including Indian. The Characterset support US ASCII, 7 Bit ISCII and 8 Bit ISCII. This is based on the standards specified for GIST Technology and Software compatable under XENIX and UNIX environments facilitating the communication through networks as effectively as English in Roman Script.

While undertaking the data processing of Indian Scripts few transliteration problems are to be tackled with care less there be any misrepresentation of Script.

eg: The Hindi Secondary consonant (Sanyuktakshar) . s to be rendered as - ------ To write the pure consonant such as 'k'. the corresponding character key followed by <d> key is to be used. Once a text is entered by this procedure in one language say Hindi, the same text can be obtained in other Indian Scripts by selecting the required language mode and recalling the text.

Another striking development in this area are the "Shrink Wrapped" packages in Indian language word processing and DTP. "Shabdaratna", 'Venus Publisher' and 'Prakashak' have carved a niche with multiple language packages.

In India, Libraries, including public and academic are maintaining literature in various Indian Languages apart from the Libraries specialising in only Indian languages should be well channelised to meet the demands of the users scattered at various places. Standardized Subject Indexing provisions are inadequate for processing Non-Roman languages as a result of this, subject indexing poses a greater problem, ultimately effecting the precision in retrieval of records/information. In this circumstances, devising a multilingual data base is very much essential to meet the expectations of UBC.

Though, there are some efforts done in the line of designing software for muiltilingual processing of Indian Language scripts the applications of the same to Library and Information Systems are not yet materialised. In view of this, it is suggested that National Library and INFLIBNET jointly make efforts at this juncture to pool all the resources like, data, software experts alongwith communication experts in consultation with Linguists in devising the multiligual software suitable to Library and Information environment.

REFERENCES

1. Wellisch, H H: International forum Inform. Doc. 3, p18-24 (1978).

2. Metcalfe, M R and Savige, D J: It's all Greek to me : Development of a Greek script catalogue (Paper presented at the 10th annual meeting of the Dobis/Libis User group, Tokyo, Japan, Sept. 1991)

3. Kent, Allen : Encyclopedia of Library and Information, v.56, p260-278 (1995)

4. Anderson, Jarome J: Communicating in a multilingual World. INICAE, V.13,no.2 (Sept. 1994).

5. BHASHA : Indian Scripts Utility and script processor. User manual. Mysore, Central Institute of Indian Languages.

6. VM4005 : Multilingual Video display terminal : User Handbook. Bangalore, VXL Instruments Limited. (1994)

7. Frontline v.12 no.25 Dec 2-15, 1995 page 110.