CALIBER-98, 4-5 March 1998, Bhubaneswar, pp.233-243

Databases in Indian Languages: Some Issues

K. S. RAGHAVAN and V. CHANDRAKUMAR
Department of Library & Information Science, University of Madras, Chennai - 600 005

Abstract

Not much progress has been made in India regarding machine readable databases containing records of materials in Indian languages in the languages and scripts of the items. Major projects that have been initiated or completed include i) machine readable version of the National Bibliography of Literature, ii) Tamil collection of the Roja Muthiah Research Library in Chennai and iii) collection of the library of the Deptt. of Indology of EFEO. Discussions regarding the use of Indian language scripts in the designing of language databases are made by citing the provisions of different codes and standards. Based on the experience gained in a project for the creation of multi-script database, the authors describe some issues related to multilingual databases, multilingual thesaurus, compatibility between different languages and provide few suggestions. There is a need for developing multilingual thesaurus in Indian languages to facilitate indexing and searching of materials available in Indian languages and scripts.

Introduction

'Databases in Indian languages' probably refer to existing and proposed databases of materials in Indian langauges. The need for creating databases in Indian langauges is well recognised. Some of the major efforts at developing bibliographies of materials in Indian languages include:

- *The Indian National Bibliography*, 1956-
- *The National Bibliography of Indian literature* / edited by B.S. Kesavan and published by the Sahitya Akedmi
- The state bibliographies and printed library catalogues brought out by many agencies, etc.

However, most of these are:

(i) Available only in print form; and

(ii) Contain records even for Indian language materials in the Roman script.

There are, however, some printed state bibliographies and library catalogues that are also available in the language and script of the item. If we interpret 'Databases in Indian languages' to mean machine readable databases containing records for materials in Indian languages in the languages and scripts of the items, it would indeed appear that there has not been much progress in this area. In fact production of machine readable bibliographic databases in India covering material in Indian languages is beginning to be considered seriously only in the recent years.

Major Projects

Some of the major projects that have been initiated or completed in recent years include:

- A project funded by U.S.based organisations and recently completed, has developed the machine readable version of the National Bibliography of Indian Literature. This is now available both on CD and also on the Web.
- A short Author-title catalogue containing over 40000 records of nearly the entire Tamil colection of the Roja Muthiah Research Library in Chennai (a project of the University of Chicago) is available in machine readable form and is accessible even on the Web.
- A database of the entire collection of the Library of the Department of Indology of EFEO (Ecole Francaise d'Extreme Orient) at Pondicherry comprising of material in Sanskrit, Tamil and other Indian languages besides English, is also available in machine readable form.

Probably there must have been many other similar developments in other parts of the country. In a recent paper, Kaul has stressed the importance of creating bibliographic databases including union catalogues of publications in Indian languages using the available technology. He also refers to efforts of DELNET in this regard[5].

Use of Indian language scripts

In recent years however, owing to a variety of factors, publishing in regional languages has been growing. Recognising this development, libraries and information institutions have tried to provide access to and display the materials in different langauges in the languages and scripts of the items. This has generally been done by creating and maintaining separate card catalogues for this purpose and this has usually been in addition to one unified catalogue using the Roman script for the entire

library collection. *The Anglo-American Cataloguing Rules, second edition, 1988 revision* in its section 1.0E1 clearly prescribes that the following areas of a bibliographic record should be transcribed from the item itself in the language and script in which it appears there[1]:

- Title and statement of responsibility
- Edition
- Publication, distribution, etc.
- Series

However, the code, as is to be expected, favours Romanization of names of persons / corporate bodies / uniform titles, etc. used as access points. The access points that are quite extensively used in information retrieval from bibliographic databases include:

- Names of persons;
- Names of corporate bodies;
- Names of subjects / keywords / descriptors
- Names of works / series; and
- Title of documents.

In considering such a recommendataion for the design of bibliographical databases of publications of Indian langauges in the context of available technology several issues arise:

- Should the heading for a person / corporate body / subject / uniform title, etc. for a publication in an Indian language be in the Roman script or in the script of the language of the item?
- Should we decide not to Romanize the heading, and the name of a person / corporate body or a uniform title appears in several Indian languages, what script should be employed to formulate headings in the bibliographic records?

As already mentioned some of these issues have been provided for by the *Anglo-American Cataloguing Rules*. The prescribed solutions are quite acceptable in an English language speaking country as it would facilitate filing of all bibliographic records in a catalogue or index terms in an inverted file in a single alphabetical sequence. The growth in regional language publishing has also made publishing industries, governments at state and national levels to initiate efforts to explore the possibility of using the new technology for regional language publishing. As a result

of these efforts, several software packages have emerged in the last decade and more. These developments have naturally resulted in increased use of new technology for publishing in Indian languages. In addition these developments have also resulted in examining the possibility of creating bibliographic records and databases using scripts of Indian languages. Among the several tools that have been developed for handling Indian scripts is the GIST Card developed by the C-DAC. GIST is compatible with the IBM PC / XT / AT computer systems, and it allows one to simultaneously use any Indian language and script along with English. GIST is compatible with commonly used word processing softwares and DBMS packages such as CDS/ISIS, dBase, Foxpro, etc.[3] This naturally means that libraries and other information institutions in India should explore the possibility of using the technology for design and development of multiscript databases in addition to concentrating on other aspects of automation. There is, therefore, a need for developing appropriate national standards in this regard. The guidelines developed by INFLIBNET for data capturing need to be extended and made more comprehensive to cover some of these issues. The GIST technology, probably for the first time, offers the facility to use Indian language scripts in bibliographic data management. The GIST technology uses extended ASCII characters to represent Indian language scripts. On the other hand, the practice in the U.S. in handling Indian language material is to Romanize the words with diacritical marks that are ignored for the purposes of filing and arrangement. The GIST files all Indian scripts in a single alphabetical sequence after all the ASCII characters. In other words the sequence would be:

A - Z

a - jña

Scope of the paper

In this paper one major issue in the design of databases in Indian languages, viz., the provision of subject access to materials in Indian languages via descriptors in Indian languages and scripts has been examined at some length. The issues raised are almost entirely based on the experience gained in a project awarded by the French Institute of Pondicherry for the creation of a multiscript database of the holdings of the collection of its Department of Indology. The following tools were employed for designing and developing the database:

- The Anglo-American Cataloguing Rules for heading and description of items;

- *The Common Communication Format* with certain modificatons for the bibliographic record format;
- An extended version of *Dewey Decimal Classificaiton, 20th edition* for assigning class numbers to the bibliographic items;
- The CDS/ISIS as the bibliographic data management software; and
- The GIST technology for handling Indian language scripts.

The database contains over 30000 records and has been in use by the olars of the French Institute of Pondicherry. While creating the abase in assigning descriptors in the CCF field 620 of each of the liographic records, it was decided to use the language and script of item. Traditionally bibliographic files have used vocabulary control ices in the English language such as the *Library of Congress List of ject Headings* (LCSH), *Sears' List of Subject Headings* (SLSH) or vocabulary of a classification scheme such as *Dewey Decimal ssification*, etc. for this purpose. In the present context it became essary to develop the facility to represent as well as provide for rching the system using descriptors in Indian languages and scripts.

ltilingual Databases

In using the GIST technology for the creation and maintenance of lti-script databases there are two options that are available:

(i) Creation and maintenance of separate databases for each of the languages and scripts using the GIST technolgy; or

(ii) Creation and maintenance of an integrated multilingual and multiscript database using the GIST technology.

Some of the major limitations of the GIST technology in maintaining integrated multiscript database are:

- The limitations in simultaneously handling and displaying records in different Indian language scripts. At present GIST has the capability of displaying Roman and only any one Indian Script. This makes it difficult to simultaneously view on a GIST screen, records in, say, Tamil and Devanagari scripts. It imposes on the viewer the limitation of having to view records in Indian languages in one script at a time. However, it does provide for automatic transliteration from any Indian script to Roman script or any other Indian script and vice-versa.
- The filing order of Indian characters built in to GIST technology is parallel to what is generally followed in Sanskrit and other

Devanagari script-based langauges. While this order is widely accepted and used in almost all the Indian languages there are some Indian languages where the resulting filing order is not the same as what is conventionally followed.

- Special characters and European language letters such as á, é, ö, ü etc. cannot be handled in the GIST environment as the ASCII codes for these characters are used for Indic characters.
- Kaul refers to another major problem in respect of publications in Urdu, Arabic, Persian, Sindhi and Kashmiri written especially in Perso-Arabic script. While we can create databases in these languages, we are not able to search through various permutations and combinations and also convert them into other langauge scripts using GIST[5].

Multilingual Thesaurus

The major problems identified above relate basically to limitations of the available GIST technology. In the context of the project referred to above, in addition to this, several conceptual issues also came up. The design of a multilingual bibliographic database with facilities to assign descriptors and input search terms in Indian languages and scripts presupposes the availability of a multilingual thesaurus or separate thesauri in the different languages and subjects of the databases. In an ideal situation the attributes and characteristics expected of such a system would include:

- The system accepting a search term in any langauge or script (including English) and allowing the user the option of automatically displaying equivalent terms in the other languages of the databases;
- Providing facilities to the searcher to select search terms in one or more languages and formulate a search;
- Providing facilities for carrying out the search in one or more databases; and

  Enabling viewing and / or displaying and / or printing the records in the scripts of the respective items or in a common script.

Compatibility issues

A major issue in the design of such an information system would be one of compatibility between the different langauges. The literature on comptibility issues defines the term compatibility as "relationship between

two systems / entities"[7]. Developments in technology have made automatic information retrieval systems an alternative means to achieve compatibility between systems. These could be particularly useful for relating terms in several separate thesauri in an online multiple database environment. Technology could also be exploited in a multilingual integrated database environment in which records for different language items are created in the language and script of the respective items and several separate langauge thesauri are used for indexing and searching. This issue is becoming especially relevant in the context of a multilingual country such as ours in which literature exists in several different languages, many of the langauges having their own script. Here we are primarily concerned with issues of compatibility relating to subject descriptors in Indian languages used for representing subjects of the documents. Literature on the subject of compatibility identifies three major types of compatibility issues that may arise in providing for switching between two or more languages[2]:

- Conceptual compatibility;
- Verbal compatibility;
- Structural compatibility;

Conceptual compatibility

Conceptual compatibility exists between two langauges if for every concept for which a term is available in one langauge, a term exists in the other langauge. The ISO 5964 refers to five types of conceptual compatibility issues that one may encounter in practice[4].

(a) *Exact equivalence*: The source langauge term is identical in meaning and scope to the target language term. The exact equivalent may be morphologically related to the term in the source language.

Examples:

Tamil	*Sanskrit*
Tattuvam	Tatvaśāstra
Tarkkam	Tarkaśāstra

They may be morphologically unrelated as in the example below:

Example:

Tamil	*Sanskrit*	*English*
Ilakkaṇam	Vyākaraṇam	Grammar

(b) *Inexact equivalence*: A term in the target language has a slightly different connotation from that of the source language term.

Examples:	*Sanskrit*	*English*
	moksa	Salvation
	nigantu	Lexicon

(c) *Partial equivalence*: A term in the source language can not be matched by an exactly equivalent term in the target language, but a nearer translation can be achieved by selecting a term with a narrower or broader meaning.

Example:	*English*	*Tamil*
	Grottos	kukaikaḷ

(d) *Single to multiple equivalence*: A term in the source language can not be matched with an exactly equivalent term in the target language, but the concept which the source language term refers to may be achieved by a combination of two or more in the target language. Three kinds of situations that have been identified:

Situation 1: A concept represented by a single term in one language may have two or more different equivalents in another language.

Example:	*Tamil*	*Sanskrit*
	tirukuṭamuzukku	mahākumbhabhiṡēkam mahāsamproksanam

Situation 2: A compound term in the source language represents a concept which is expressed using two or more terms in the target language.

Examples:	*English*	*Tamil*
	Mythology	purāna ilakkiyam
	Public fast	nōṉpu

Situation 3: A term in the source language refers to an extra category which is not evolved for cultural or linguistic reasons, in the target language.

Examples:	*Tamil*	*Sanskrit*	*English*
	tūtu	?	?
	Piḷḷaitamizh	?	?

(e) *Non-equivalence*: The target language does not contain a term which corresponds in meaning even partially or inexactly to the source language term.

Examples:	*Sanskrit*	*Tamil*	*English*
	agama	?	?
	sandhi	canti	?

Verbal compatibility

In languages which contain many words often derived from the same root (such as Sanskrit or the root Dravidian language), verbal compatibility exists if one term that is in use in two or more language denotes the same concept in all the languages. The problems of incompatibility that may arise are:

(a) Inter-language homographs: When two identical words in two different languages denote different concepts a problem of compatibility arises:

Example:	*Tamil*	*Malayalam*
	vellam (Flood)	vellam (Water)

Structural compatibility

In the present context the term structure is used to refer to the semantic sturucture of a term. Traditionally thesauri and descriptor languages have recognised hierarchically (BTs and NTs) and non-hierarchically (RTs) related terms. The hierarchy of a set of concepts (terms) in one language may not be amenable for exact mapping / matching on to another language.

Example:

Tamil	*English*
pāl	Gender
NT āṇpāl	NT Masculine gender
NT peṇpāl	NT Feminine Gender
NT palarpāl	?
NT oṉṟaṉpāl	?
NT paviṉpāl	?

Conclusions and Suggestions

The major limitations of GIST technology in this regard appear to relate to the use of 8-bit codes for representing characters in today's computers. This problem of character set in bibliographic data exchange has not been given in the same degree of attention and consideration as certain other areas of automation such as bibliographic record formats, standards for bibliographic description (ISBDs), etc. The available technology of characters encoding does not fully meet the requirements of multilingual multiscript envrionment such as the one obtaining in India. There has been a discussion and a debate over the need for moving to a 16-bit or 32-bit character sets to accommodate all world scripts in current use. The Unicode consortium formed in 1991 and the ISO initiative in 1983 to develop a new standard for character encoding are some of the major international initiatives in this regard[6]. The version 1.01 of Unicode which was published includes some 27000 characters. The development of a 16-bit code appears to be the only probable solution, although the price of such a conversion will need to take into consideration the considerable stock of hardware already installed around the world.

The subject of developing databases of material in Indian languages in the langauge and script of the item is an important problem for consideration by agencies such as INFLIBNET, Bureau of Indian Standards, etc. in developing appropriate guidelines and standards in this regard.

The need for creating multilingual databases is now fairly well recognised. There is therefore a need for developing multilingual thesaurus in Indian languages to facilitate indexing and searching of materials available in Indian languages and scripts.

References

1. *Anglo-American Cataloguing Rules. 2nd edition, 1988 revision*; 1988; p. 15.
2. Dahlberg, I. Towards establishment of compatibility between indexing languages. *International Classification.* 1981; 8 (2); 88.
3. GIST: Multilingual card, user's guide. Kanpur; Quark Computers; 1994; pp.2.3.
4. ISO 5964. *Documentation: Guidelines for the establishment and development of multilingual thesaurus.* 1985; pp.11.

Conclusions and Suggestions

The major limitations of GIST technology in this regard appear to relate to the use of 8-bit codes for representing characters in today's computers. This problem of character set in bibliographic data exchange has not been given in the same degree of attention and consideration as certain other areas of automation such as bibliographic record formats, standards for bibliographic description (ISBDs), etc. The available technology of characters encoding does not fully meet the requirements of multilingual multiscript envrionment such as the one obtaining in India. There has been a discussion and a debate over the need for moving to a 16-bit or 32-bit character sets to accommodate all world scripts in current use. The Unicode consortium formed in 1991 and the ISO initiative in 1983 to develop a new standard for character encoding are some of the major international initiatives in this regard[6]. The version 1.01 of Unicode which was published includes some 27000 characters. The development of a 16-bit code appears to be the only probable solution, although the price of such a conversion will need to take into consideration the considerable stock of hardware already installed around the world.

The subject of developing databases of material in Indian languages in the langauge and script of the item is an important problem for consideration by agencies such as INFLIBNET, Bureau of Indian Standards, etc. in developing appropriate guidelines and standards in this regard.

The need for creating multilingual databases is now fairly well recognised. There is therefore a need for developing multilingual thesaurus in Indian languages to facilitate indexing and searching of materials available in Indian languages and scripts.

References

1. *Anglo-American Cataloguing Rules. 2nd edition, 1988 revision*; 1988; p. 15.
2. Dahlberg, I. Towards establishment of compatibility between indexing languages. *International Classification.* 1981; 8 (2); 88.
3. GIST: Multilingual card, user's guide. Kanpur; Quark Computers; 1994; pp.2.3.
4. ISO 5964. *Documentation: Guidelines for the establishment and development of multilingual thesaurus.* 1985; pp.11.

5. Kaul, H.K. From printed bibliographies to online databases : role of library networks. *University News*. June 1997; 35 (24); 13.

6. Peruginelli, Susanna, Bergamin, Giovanni and Ammendaola, Pino. Character sets : towards a standard solution? *Program*. July 1992; 26 (3); 218.

7. Unesco. UNISIST Study Report on the Feasibility of a World Science Informations System. Paris; Unesco; 1971; p. 97.