MULTILINGUAL ACCESS TO INFORMATION IN A NETWORKED ENVIRONMENT CHARACTER ENCODING & UNICODE STANDARD

Subal Chandra Biswas

Abstract

With the recent rapid dissemination over the international computer networks of world-wide distributed document bases, the question of multilingual access and multilingual information retrieval is becoming increasingly relevant. Briefly discusses some of the issues that must be addressed in order to implement a multilingual interface for a digital library system in general, and the problems associated with character encoding of multilingual text in particular. Highlights the development of Unicode Standard, which is envisioned as a solution to the problem of multilingual character encoding and its progressive implementation by the computer industry, including the library community. Concludes that library and information networks in India must adopt Unicode to digitize and provide access to the rich repertoire of literature generated in scores of Indian languages.

Keywords: Multilingual digital information retrieval; Character encoding; Unicode Standard.

1. INTRODUCTION

One of the great benefits of digital libraries is the ability to make information available to a wide audience without geographic constraints. As large-scale digital libraries contribute to the dismantling of the geographic barriers to information access, however, the barrier raised by language considerations becomes apparent. In order to make information available to the widest possible audience, this language barrier must also be tackled [Sheridan 1997]. Ideally, a global digital library would provide access to information in all the languages of the world, to all people, all the time. Anyone anywhere could create information in their native language, yet others, wherever they are, could discover that information and have it translated into their favoured languages and formats. Such an ideal situation may well be beyond our reach at the present moment, but it raises issues worthy of investigation.

No single collection, user interface, or set of system capabilities will serve young and old, novice and expert, artist and physicist, in New Delhi and New York. Nor will any single system provide adequate access to books, music, movies, and numeric data, much less serve applications as diverse as e-commerce, weather modelling, census tracking, library catalogues, and virtual classrooms. Yet people with varying backgrounds and skills, speaking in different tongues, have similar information needs. Content of interest may exist in a wide variety of forms and languages, and the same content and collections may be of interest to a wide range of people, for many different reasons. The prospect of a global digital library presents several opportunities like,

- Making information resources accessible to particular user communities while at the same time making those same resources accessible to a broader, improperly-defined, and perhaps unknown audience, and
- Enabling users to bridge the many formats, representations, and languages of individual digital libraries in their quest for information resources.

The issue of multilingual access to information is an example of the tradeoffs between tailoring to local communities and generalising to a global audience. This particular issue is

1. Urgent in view of the rate at which new text is being created,
2. Long-term — in view of the relative stability of languages compared to the ephemeral nature of new media formats, and
3. Broad-reaching — in view of the fact that some form of textual content is involved in most forms of electronic communication, whether alone or as descriptions of other media.

This paper addresses issues related to multilingual access to digital information repositories in general, and the problems associated with character encoding of multilingual text in particular with special emphasis to the practices followed by library community in this regard.

2. FROM MONOLINGUAL TO MULTILINGUAL DIGITAL LIBRARIES

Language is one of the most critical factors in access to information. Information is usable and useful only if it is in a language that can be read or otherwise understood. Communication between people speaking a common language, whether English, French, Mandarin, or Hindi, is influenced by factors such as subtleties of phrasing, vocabulary, regional accents, and knowledge of the subject matter being discussed. In monolingual information retrieval, some of these difficulties are ameliorated through

- Standardising forms (e.g., singular/plural, stemming to remove variant word endings),
- Controlled vocabulary (e.g., thesauri that establish preferred terms and make cross-references from synonyms and related terms), and
- Algorithms that employ knowledge of grammatical structure and word frequency in the language.

People often need information that exists in languages other than those they read or speak, however. Human translators can bridge language in oral and written communication. They can translate between spoken languages in real time (as interpreters in international meetings, conferences, etc.), and they can translate texts. Translation is a challenging intellectual task, as no two languages map precisely onto one another. A word in one language may have several meanings in the other, and the interpreter must determine which is most appropriate. Conversely, one word may incorporate multiple meanings in another language and the appropriate subset must be identified. Often, a narrative explanation is needed in place of a single word. Meaning depends on context as well as on choice of words. Metaphors and idioms are particularly difficult to convey in other languages. Translators are really interpreters of meaning; that is why automatic translation is so limited. Computers can provide approximate translations, particularly of scientific texts, but are far from a reliable substitute for expert human language interpreters.

Providing access to information in multiple languages is a challenge inherent in constructing a global information infrastructure. It is also fundamental to tailoring and interoperability trade-offs. Given a choice, people generally prefer to communicate in their native language, both online and offline. Communication in international environments frequently involves multiple languages, however, and people often need information written in unfamiliar languages. The problem of multilingual access to information affects not only users in countries in which several languages are spoken (like India), but also all those who search material in information sources or databases containing material in more than one language. This is the case in the majority of scientific or research databases (like INSPEC, Medline, ERIC, etc.). In addition, the growth of networks means that we can easily access information/databases outside our own immediate circle — in another country, another continent. In doing so, we encounter problems concerning not only search interfaces, but also concerning search keys like, subject or even author in another language [Lehtinen and Clavel-Merrin 1998]. Multilingual information access is a pervasive problem in automation, and it is of great concern for anyone exchanging information over computer networks. Multilingual access issues affect e-commerce, information institutions such as

libraries, archives, museums, academic institutions at all levels, and those who produce hardware and software for network applications.

However, it is important to understand the connotation of the term 'multilingual access.' Multilingual access is frequently used to cover a variety of topics, and it is necessary to define the use of the term 'multilingual' in the context of this paper. We frequently hear talk of multilingual systems in reference to search environments (e.g., display screens, user dialogue, help screens, etc.), whereas the actual access points themselves are in fact monolingual. Multilingual user environments are now standard in the majority of systems.

3. MULTILINGUAL ACCESS TO INFORMATION

Multilingual digital libraries are being created by countries with more than one national language, by the European Union, and by international research consortia and international businesses. Major scientific, technical and medical databases, such as INSPEC and Medline, long have contained metadata in English to represent materials in other languages. Monolingual digital libraries also are being created in many different languages. Many of these digital libraries are of interest well beyond the borders of the countries creating them, and to users who are native speakers of other languages. A fundamental challenge of constructing a global information infrastructure and a global digital library is to provide access to these collections, regardless of the language of the content and the language of the information seeker.

The World Wide Web offers access to information resources in many languages. Certain developments facilitate multilingual exploitation of these resources. Some search engines, for example, allow the user to restrict retrieved sites to those in particular languages; some also provide the searcher with an interface in a chosen language. Many Web sites also offer their information in several languages, one of which typically is English. Systran, a machine translation system available from the AltaVista search engine, can even translate a search statement or a retrieved page from one language to another. Despite these features, however, language also creates obstacles to full exploitation of Web resources. Not all languages are catered for by these multilingual tools. Machine translation output typically is but a rough and ready version of a human translation. The varieties of scripts in which the written forms of the world's languages appear also create major problems in searching, inputting, displaying and printing text in non-Roman scripts [Large and Moukdad 2000].

Multilingual access to information is a complex problem. Peters and Picchi [1997] have divided it into two basic parts:

1. Multiple-language recognition, manipulation, and display — involving matters such as encoding character sets and symbols so that they can be manipulated (sorted, searched, or otherwise exploited), and
2. Multilingual or cross-lingual search and retrieval — referring to search for content in other languages, otherwise known as "cross-language information retrieval."

The first point addresses the problem of allowing digital library users to access the system, no matter where they are located, and no matter in what language the information is stored; it is a question of providing the enabling technology.

The second point implies permitting the users of a digital library containing documents in different languages to specify their information needs in their preferred language while retrieving documents matching their query in whatever language the document is stored; this is an area in which much research is now under way. Let us first have an outline of this second point

3.1 Cross-Language Information Retrieval

As explained in the previous section, the interface of a multilingual digital library must include features to support all the languages that will be maintained by the system and to permit easy access to all the documents contained. However, it must also include functionalities for multilingual or cross-language search and retrieval. This implies the development of tools that allow users to interact with the system, formulating the queries in one language and retrieving documents in others. The problem is to find methods which successfully match queries against documents over languages. This involves a relatively new discipline, generally known as Cross-Language Information Retrieval (CLIR), in which methodologies and tools developed for Natural Language Processing (NLP) are being integrated with techniques and results coming from the Information Retrieval (IR) field.

Peters and Picchi [1997] have identified three approaches that are being studied for solution of the CLIR problem:

1. Text translation by machine,
2. Knowledge-based techniques, and
3. Corpus-based techniques.

Text translation can be divided further:

a) Translating the full content of the digital library into another language, or

b) Translating only the query [Oard 1997; Peters and Braschler 2001].

Translating the full content is rarely feasible except in very small and specific applications. For example, in the area of meteorological reports automatic translation is reasonably successful. For most applications, translating queries appears to be more promising. Queries in English, say for example, can be translated into Japanese and searched in Japanese databases, and vice versa.

Knowledge-based approaches involve multilingual dictionaries, thesauri, or ontologies [Peters and Picchi 1997; Clavel-Merrin 1997]. In these approaches, searchers can construct a query using terms in their own language. The multilingual thesaurus is then used to translate the terms into the target language; they are then submitted to databases in that language. Thesauri can be translated pair wise (say, between English and French) or among a larger number of languages (say, English, French, German, Spanish, and Russian). In the latter case, a common core of terms would be established, and then the linguistic equivalents in each language would be determined [Soergel 1997].

Corpus-based techniques involve linguistic characteristics of a language and the distribution of terms in a collection of documents. Information retrieval techniques developed for monolingual retrieval, such as vector-space models and probabilistic methods can be employed for multiple languages [Bian and Chen 2000; Oard 1997; Peters and Picchi 1997]. Usually some test databases in each language are required to “train” the algorithms about the relationships between corpora (bodies of text) in each language.

Each of these methods is both promising and problematic. None are complete solutions, and all are under active study. For any of the cross-language techniques to be effective, however, underlying technical issues in managing multilingual text must be resolved.

3.2 Multilingual Recognition and Representation

Despite its name, until recently the World Wide Web had not addressed one of the basic challenges to global communication: the multiplicity of languages. Standards for protocols and document formats originally paid little attention to issues such as character encoding, multilingual documents, or the specific requirements of particular languages and scripts. Consequently, the vast majority of WWW browsers still do not support multilingual data representation and recognition. Ad-hoc local solutions currently abound which, if left unchecked, could lead to groups of users working in incompatible isolation.

The main requirements of a multilingual application are to:

1. Support the character sets and encodings used to represent the information being manipulated;
2. Present the data meaningfully;
3. Manipulate multilingual data internally [Peters and Picchi 1997].

For the rest of this paper I shall concentrate on the first requirement of multilingual application, i.e. encoding of character sets in various languages.

4. CHARACTER ENCODING

Character encoding is at the heart of the data-recognition, data-manipulation, and display components of multilingual information access. The creation of characters in electronic form involves hardware and software to support input, storage, processing, sorting, displaying, and printing. Each character – or ideograph in languages such as Chinese, Japanese, and Korean – needs a unique code to ensure that it sorts, displays, prints, and is matched properly upon searching. Additional codes are required for punctuation, direction of text (left to right or right to left), carriage returns, and line feeds. The internal representation of each character determines how it is treated by the hardware (keyboard, printer, display, etc.) and the application software (sorting, searching, etc.).

To the dismay of those whose languages have far larger character sets (e.g. Devanagari with 48 alphabets, Russian with 32 alphabets; and not to mention the 2000 Kanji character Japanese), typewriter keyboards were initially designed for the English language; thus, they contain only 26 letters, 10 numbers, basic punctuation, and a few special symbols (like, @, $, &, etc.). Only a few more symbols and function keys were added when the typewriter keyboard (popularly known as the QWERTY keyboard) was adapted to become a computer keyboard. Many versions of the QWERTY keyboard are now in use worldwide, varying the location of letters. Special characters, such as letters with diacritics, can be generated by programming individual keys or by programming key sequences, often using escape key. The same key sequences on two different keyboards may produce two different characters, depending on the encoding system employed. Conversely, key sequences required to generate a character vary by keyboard and encoding system.

Effective data exchange is heavily dependent on character encoding. Characters produced from different applications may appear the same on a screen or a printout but have different internal representations. Merging or transferring data between applications requires a common character-encoding standard or software to map variant encoding formats. Searchers need to have the appropriate keyboards and software to generate characters in the encoding standard in which the contents of a digital library are stored, whether locally resident or available through mapping software located at the digital library site or elsewhere. Local printers and displays must have the appropriate software to interpret and produce characters accurately. These factors are not specific to digital libraries; rather, they are issues for all distributed applications on a global information infrastructure.

5. MONOLINGUAL, MULTILINGUAL, AND UNIVERSAL CHARACTER SETS

For character encoding many standards and practices exist, viz.

- Those that are language specific,
- Those that are script specific (e.g. Latin or Roman, Arabic, Cyrillic, Hebrew), and
- The remaining are universal standards that support most of the world's written languages.

Digital libraries employ many different character-encoding formats, and this often leads to problems in exchanging data.

After many years of international discussion on the topic, it became evident that adopting a universal character set offered the greatest hope for exchanging text in digital form. If data in all written languages were encoded in a common format, then data could be exchanged between monolingual and multilingual applications, whether email, e-commerce, or digital libraries. Which common format to accept was a matter of long debate, however. In 1991, the Unicode Consortium and the International Organization for Standardization (ISO) finally reached an agreement to merge the 16-bit Unicode Standard and the 32-bit ISO standard into a common 16-bit Unicode Standard. Version 1.1 was first published in 1993; version 4.0 is currently the accepted Unicode Standard (ISO/IEC 10646) and the latest revision is numbered as 4.1.0 [Erickson 1997; The Unicode Consortium 2003-2005].

Unicode can support more than 96,000 distinct characters. Version 4.0 of the standard provides about 96,382 characters from the world's alphabets, ideograph sets, and symbol collections. Version 4.1.0 of the standard has included a further 1273 new characters [The Unicode Consortium 2003-2005]. A growing number of world's major hardware and software vendors are supporting it and it is being incorporated into popular operating systems and programming languages. A list of such Unicode enabled products can be found from the Unicode Consortium's homepage. As software for e-commerce, digital libraries, automated library management systems, and other applications begin to support Unicode, it will become more widely adopted [Tull 2002; Needleman 2000]. Unicode Standard represents a tremendous advance for multilingual computing, but more work is needed in order to achieve a truly multilingual Web. The impact will be manifold: it will help in the publication and preservation of materials in ancient and historic scripts, facilitate the online teaching of languages using such scripts, and provide universal access to our cultural and literary heritage [Anderson 2003]. As storage costs continue to decline, the storage requirements of Unicode will be an insignificant issue, particularly for new applications.

In the meantime, substantial amount of text continue to be produced not only in language-specific and script-specific encoding standards but also in local and proprietary formats. Any of this text maintained in digital libraries may become legacy data that has to be mapped to Unicode or some other universal standard in the future. For the present, digital library designers face difficult trade-offs between the character-set standards in use by current exchange partners and the standard likely to be in worldwide use in the future for a broader range of applications.

6. TRANSLITERATION AND OTHER FORMS OF DATA LOSS

A long-established intermediate approach to character encoding for languages that cannot be typed on a standard computer keyboard is transliteration, which matches characters or sounds from one language to another but does not translate meaning. Languages written in non-Roman scripts, such as Japanese, Arabic, Chinese, Korean, Persian, Hebrew, and Yiddish (the "JACKPHY" languages), Russian, and Devanagari, are transliterated into Roman characters in many applications. Transliteration is necessary when mechanical devices such as typewriters and computers do not support the necessary scripts. It

also is helpful for people without full proficiency in the scripts of the language (e.g. recognition of Chinese or Russian names or terms transliterated in English-language contexts, such as “Beijing” or “Gorbachev”). The transliteration process may be irreversible, and thus some data loss takes place. Despite the existence of an international standardization body (ISO/TC46/SC2: Conversion of Written Languages), multiple transliteration systems exist for Cyrillic, Asian, and other character sets. Thus transliteration can be inconsistent, which is why the same Russian name may appear as “Tchaikovsky” or as “Chaikovoskii” depending on which system is used to transliterate Cyrillic characters.

Data also are lost when languages written in extensions of the Roman character set, such as French, Spanish, German, Hungarian, Czech, and Polish, are stored without the diacritics (accents, umlauts, and other language-specific marks) that for additional characters (e.g., ó, ò, ô, ö, õ, and o are distinct characters), all of which are collapsed into o if the diacritics are omitted).

Data loss is not exclusive to text, however; it is widespread in visual and aural media. Image compression is essential for efficient transmission of either still or moving pictures, and application-specific image-compression standards exist for such applications as facsimile transmission (or fax), pictures, moving images, and high-definition television. Compression algorithms are made more efficient by discarding some of the data to reduce the granularity of the image. Images subjected to “lossy” compression are legible and suitable for many applications, but the original image cannot be reconstructed. “Lossless” compression retains the full content but requires more storage and more transmission time [Harold 1996]. Similarly, for music, a popular compression algorithm known as MP3 is used to squeeze digital audio by a ratio of 12:1. MP3 compresses music sufficiently that audio files can be sent via the Internet or stored on a hard disk. Music compressed in MP3 can be expanded and played by an MP3 player with a near-CD-quality sound, yet the files are small enough to be attached to an email message. The music industry is promoting another standard, and Microsoft is promoting yet a third standard named Windows Media Audio for music compression and distribution [Audio file format 2005]. The acceptance of technical standards for music and other audio may be determined as much by the availability of playback devices as by the quality of reproduction.

The amount of data loss that is tolerable varies by application. Far more data loss is tolerable in email and in teleconferencing (where speedy communication tends to be valued over authoritative form) than in financial, legal, or bibliographic records (where validation is essential). Textual data that have been transliterated or stripped of diacritics are likely to contain variant forms of words that will not match and sort properly, incomplete words that will not exchange properly with digital libraries using complete forms, and incomplete forms that are not adequate for authoritative or archival purposes.. In digital libraries, any kind of data loss can result in information retrieval errors (e.g., items that cannot be located).

7. PRACTICES OF THE LIBRARY COMMUNITY

An essential component of any library application is an encoding methodology that allows computers to process characters and symbols used to represent language information in written form. For years the encoding mechanism was not developed under a unified umbrella nor did it reach various languages equally. Without a standard unified character code, users have to use different software and terminals to display or enter data in different languages, especially when dealing with more than a few scripts [Barry 1997; Zhang and Zeng 1999].

The international library community began developing large multilingual digital libraries in the 1990s. Language representation is a particular concern for major research libraries, whose collections may include materials in as many as 400 written languages [LeVan 2000]. Standards for record structure and character sets were established long before either Unicode or the Internet was created. Hundreds of millions of bibliographic records exist around the world in variations of the MARC format, although in

multiple character-set-encoding formats and multiple forms of transliteration. USMARC was implemented with the American Library Association (ALA) character set, which extends the English language keyboard to include diacritics from major Roman-script languages [Fayen 1989]. The ALA character set is used by Online Computer Library Center (OCLC), the largest bibliographic utility in the world, and by most American library applications; languages not included in the ALA character set tend to be transliterated. The increasing interest in adapting USMARC and other forms of MARC to Unicode reflects the fact that other systems (present and future) will have to accommodate the huge quantity of MARC records that already exist in other character-set formats [Aliprand 2000; Tull 2002].

The Library of Congress – which contributes its records in digital form to OCLC, to RLG (the Research Libraries Group, erstwhile RLIN or Research Libraries Information Network, the other major US-based bibliographic utility), and to other cooperatives in Europe and elsewhere – has done original-script cataloguing for the JACKPHY languages since the early 1980s. RLG pioneered the ability to encode the JACKPHY languages in their original-script form for bibliographic records, using available script specific standards [Aliprand 1992]. The Library of Congress, OCLC, RLG, and other bibliographic utilities around the world exchange records encoded in full-script form.

Libraries took a long-term view of their data, capturing and representing non-Roman characters in their fullest form many years before search and display technologies were widely available [Wellisch 1978]. Full vernacular script in these records can be printed on catalogue cards, but until very recently scripts could only be displayed on computers with special equipment. Even though JACKPHY records from OCLC, RLG, and other sources are loaded into the online catalogues of individual libraries, most applications only support searching the transliterated forms. Integrated library automation software that supports Unicode is becoming more widely available, thus allowing records in non-Roman scripts to be displayed on most terminals and computers. The Library of Congress has implemented a new integrated library system that incorporates Unicode, enabling them to display their JACKPHY records on public terminals in full form. Previously, these records were viewable only through other systems outside the Library of Congress.

The international library community is more and more concerned about multilingual, multi-script data exchange as new regions of the world come online. The European Union is promoting Unicode and is funding projects to support Unicode implementation in library automation [Peruginelli et al. 1992]. In the mid 1990s, Borgman [1996] conducted a study of six countries in Central and Eastern Europe (viz. Croatia, the Czech Republic, Hungary, Poland, Slovakia, and Slovenia), each with its own national language and character set, found that a variety of coding systems were in use. More than half of the research libraries in the study used Latin2, one used Unicode, and the rest used a national or a system-specific format to encode data; none used the ALA character set.

8. CONCLUSION

As libraries, archives, museums, and other cultural institutions throughout the world become increasingly aware of the need to preserve digital data in archival forms, character-set representation becomes a political as well as a technical issue. Many agencies are supporting projects to ensure preservation of bibliographic data in digital forms that can be readily exchanged, including the Commission of the European Communities and the International Federation of Library Associations and Institutions (IFLA) [Andreoni et al. 1999].

This matter also concerns India, a country with rich diversity in languages, cultures, customs and religions, which are stored in print media as well as in numerous manuscripts, tamra-patras (copper plates), palm leaves, etc. The number of languages recorded is 418, out of which 408 are living and the remaining 11 are extinct. Eighteen are constituionally recognised languages and written in a variety of scripts. This

awesome diverisity of languages has imposed tough hurdles on the way of digitization and access to Indian literature. For a long time, romanization and use of transliteration were viewed as the only solution. But, Prasad [2003] has rightly observed that “Instead of imposing one language and one script, ideally all the languages and scripts are to be given importance, as each language has produced rich literatire and scientific information over the centuries”. Library and information networks in India hold the responsibility to digitize all those valuable resources stored in print and other media and make them accessible to users through the Web. However, due to technology limitations, so far it has not been practical to do so. Thanks to technologies like Unicode, most computers in future will support Indian scripts as well [Chandrakar 2003; Sadagopan 2000].

In this paper, I have tried to focus on only one mode of textual information storage and retrieval, viz. the character-coded text. Considerable research is underway on access to multilingual information in formats other than character-coded text, such as handling of multilingual speech, document images, and video OCR (optical character recognition) [Oard et al. 1999]. We also know that information can also be gleaned from audio and image (both still and moving) documents. Inclusion of these introduces further complication in the information retrieval scenario. However, interest in provision of access to multilingual and multimedia information in the networked environment and beliefs of professionals and users to reap the benefits of a truly global information infrastructure for betterment of our present and future civilizations are bringing together researchers from academic, research and government establishments of one nation or group of nations or whole world under single umbrellas. U.S. National Science Foundation and the European Commission are funding such projects for examining multilingual information access and management from the perspectives of digital libraries and computational linguistics [Klavans and Schäuble 1998; Hovy et al. 1999]. Similarly, presentations on aspects of the problem now routinely appear at major national and international conferences on digital libraries, information retrieval, machine translation and computational linguistics.

9. REFERENCES

1. Aliprand, J.M.: Nonroman scripts in the bibliographic environment. Information Technology and Libraries, 11 (2), June 1992, 105-119.
2. Aliprand, Joan M.: The Unicode Standard: its scope, design principles, and prospects for international cataloging. Library Resources and Technical Services, 44 (3), July 2000, 160-167.
3. Anderson, Deborah: Unicode and historic scripts. Ariadne, (37), Oct 2003-Dec 2003 <http:// www.ariadne.ac.uk/issue37/anderson/>
4. Andreoni, A., Baldacci, M.B., Biagioni, S., Carlesi, C., Pagano, P., Peters, C. and Pisani, S.: The ERCIM Technical Reference Digital Library. Meeting the requirements of a European community within an international federation. D-Lib Magazine, 5 (12), December 1999. <http://www.dlib.org/ dlib/december99/peters/12peters.html>
5. Audio file format. In: Wikipedia, the free encyclopedia. May 24, 2005. <http://en.wikipedia.org/ wiki/ Audio_file_format>
6. Barry, R. K.: The role of character sets in library automation: the development of 8 bit sets and Unicode. International Cataloguing and Bibliographic Control, 26 (1), Jan/Mar 1997, 14-17.
7. Bian, G.-W. and Chen, H.-H.: Cross-language information access to multilingual collections on the Internet. Journal of the American Society for Information Science, 51 (3), February 2000, 281-296.

8. Borgman, Christine L.: Automation is the answer, but what is the question? Progress and prospects for Central and Eastern European libraries. Journal of Documentation, 52 (3), September 1996, 252-295.

9. Chandrakar, R.: Multi-script bibliographic database: an Indian perspective. Online Information Review, 26 (4), 2002, 246-251.

10. Clavel-Merrin, G.: Multilingual access to libraries' databases. In: Towards a worldwide library: a ten year forecast. Proceedings of the Nineteenth International Essen Symposium 23-26 September 1996, edited by Ahmed H. Helal and Joachim W. Weiss. Essen, Germany: Universitatsbibliothek Essen, 1997; pp.168-79.

11. Erickson, Janet C.: Options for presentation of multilingual text: use of the Unicode Standard. Library Hi Tech, 15 (3-4), 1997, 172-188.

12. Fayen, Emily Gallup: The ALA character set and other solutions for processing the world's information. Library Technology Reports, 25 (2), March-April 1989, 253-273.

13. Harold, Thomas G.: Common graphics formats. <http://www.dreamartists.com/oldsite/gfxfmts. htm>.

14. Hovy, E., Ide, N., Frederking, R., Mariani, J. and Zampolli, A. (eds.): Multilingual Information Management: Current Levels and Future Abilities. 1999. <http://www.cs.cmu. edu/People/ref/mlim/

15. Klavans, J. and Schäuble, P.: Summary review of the Working Group on Multilingual Information Access. In: Report of the Joint US National Science Foundation-European Union Working Groups on Future Developments for Digital Library Research. ERCIM Technical Report, No.98/W004, 1998 <http://www.iei.pi.cnr.it/DELOS/NSF/Brussrep. htm>

16. Large, A. and Moukdad, H.: Multilingual access to web resources: an overview. Program, 34 (1), 2000, 43-58.

17. Lehtinen, Riitta and Clavel-Merrin, Genevive: Multilingual and multi-character set data in library systems and networks: experiences and perspectives from Switzerland and Finland. In: Multi-script, Multilingual, Multi-character Issue for the Online Environment: Proceedings of a Workshop Sponsored by the IFLA Section on Cataloguing, Istanbul, Turkey, August 24, 1995, edited by John D. Byrum, Jr. and Olivia Madison. München: K.G. Saur, 1998; pp. 67-91. (IFLA publications; 85).

18. LeVan, R.: OCLC continues systems work to support international data. OCLC Newsletter, (243), Jan/Feb 2000, 23-26.

19. Needleman, M.: The Unicode Standard. Serials Review, 26 (2), 2000, 51-54.

20. Oard, D.: Serving users in many languages: cross-language information retrieval for digital libraries. D-Lib Magazine, 3 (12), December 1997. <http://www.dlib.org/dlib/december97/ 12oard.html>

21. Oard, D., Peters, C., Ruiz, M., Frederking, R., Klavans, J. and Sheridan, P.: Multilingual Information Discovery and AccesS (MIDAS). A Joint ACM DL'99/ACM SIGIR'99 Workshop. D-Lib Magazine, 5 (10), October 1999. <http://www.dlib.org/dlib/ october99/10oard.html>

22. Peruginelli, Susanna, Bergamin, Giovanni and Ammendola, Pino: Character sets: towards a standard solution? Program, 26 (3), July 1992, 215-223.

23. Peters, C. and Braschler, M.: Cross-language system evaluation: the CLEF campaigns. Journal of the American Society for Information Science and Technology, 52 (12), Oct 2001, 1067-1072.

24. Peters, C. and Picchi, E.: Across languages, across cultures: issues in multilinguality and digital libraries. D-Lib Magazine, 3 (5), May 1997 <http://www.dlib.org/dlib/may97/peters/05peters. html>

25. Prasad, A.R.D.: Unicode: a tutorial. In: Library and Information Networking: NACLIN 2003, edited by H.K. Kaul and B.B. Das. New Delhi: DELNET, 2003; pp. 432-448.

26. Sadagopan, S.: Libraries in the dot.com era. SRELS Journal of Information Management, 37 (1), Mar 2000, 1-4.

27. Sheridan, P.: Introduction. In: Third DELOS Workshop: Cross-Language Information Retrieval, Zurich, 5-7 March 1997. ERCIM Workshop Proceedings - No. 97-W003. <http://www.ercim. org/ publication/ws-proceedings/DELOS3/#anchor552645>

28. Soergel, D.: Multilingual thesauri in cross-language text and speech retrieval. In: Working Notes of AAAI Spring Symposium on Cross-Language Text and Speech Retrieval. Stanford, CA, 1997, 164-170. <http://www.ee.umd.edu//medtab/filter/sss/ papers/>

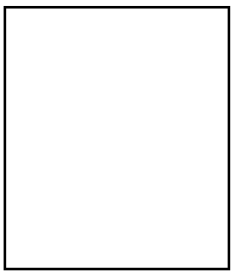
29. Tull, Laura: Library systems and Unicode: a review of the current state of development. Information Technology and Libraries, 21 (4), Dec 2002, 181-185.

30. The Unicode Consortium: The Unicode Standard, Version 4.1.0, defined by: The Unicode Standard, Version 4.0 (Boston, MA, Addison-Wesley, 2003. ISBN 0-321-18578-1), as amended by Unicode 4.0.1 (2004) <http://www.unicode.org/versions/ Unicode4.0.1> and by Unicode 4.1.0 (2005) <http://www.unicode.org/versions/Unicode 4.1.0>

31. Wellisch, Hans H.: Multiscript and multilingual bibliographic control alternatives to Romanization. Library Resources and Technical Services, 22 (2), 1978, 179-190.

31. Zhang, Foster J. and Zeng, Marcia Lei: Multiscript information processing on crossroads: demands for shifting from diverse character code sets to the Unicode Standard in library applications. IFLA Journal, 25 (3), 1999, 162-167.

About Author



Dr. Subal Chandra Viswas has hold master degrees in Economics and LIS as well as PhD in LIS. I worked as a Commonwealth Research Scholar in the Dept. of Library & Information Studies, Loughborough University of Technology, UK, during 1985-1989. Besides my current position, I have served on the faculty of Jadavpur University, Kalyani University, University of California, Los Angeles, and Banaras Hindu University. The areas of my research interest are: linguistic and cognitive bases of information retrieval, indexing languages & systems, community information, etc. I have over 50 research papers published in international and national journals, conference proceedings, composite books, etc.