DESIGNING A DIGITAL LIBRARY WITH BENGALI LANGUAGE S UPPORT USING UNICODE

Rajesh Das **Biswajit Das** **Subhendu Kar** **Swarnali Chatterjee**

Abstract

Unicode is a 32-bit code for character representation in a computer. It is described that how to use Unicode in digital library to build a Institutional Repository. Unicode Consortium develops Unicode. It is very helpful that creation of multilingual language database.

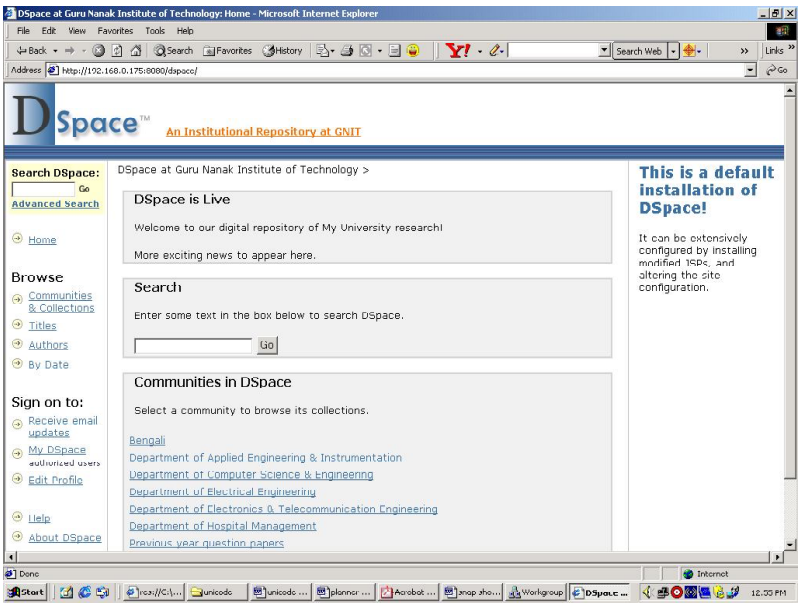
Keywords : Digital Library, Unicode, Bengali Language

1. INTRODUCTION

Digital revolution has entirely altered our lifestyle. Libraries and Information centers and their professionals are not exception to this. The information technology has altered the mode of publication in such a way that the sources of information are flooded in an attractive digital form of publications. In the changing scenario, libraries and librarians will have to play a crucial role in managing these digital resources with different language and script. Digital Library software plays an important role in organizing these digital resources with various languages. An experiment was done with the help of DSpace Digital Library Software for creating such digital collections with Bengali language using Unicode. Unicode is 16-bit code for character representation in a computer. Unicode is designed by Unicode consortium. It represents almost all the world script extinct many of extinct scripts like Bramhi and Kharosthi. ISCII is another code developed for to represent the Indian characters in computer but there are problems with character representation using ISCII. It is found that Unicode can solve the problem. This paper suggests measures for creation of Digital library in Bengali languages and the problem associated with Unicode.

2. DIGITAL LIBRARY AND DSPACE

A digital library is an integrated set of services for capturing, cataloging, storing, searching, protecting, and retrieving information, which provide coherent organization and convenient access to typically large amounts of digital information. Digital libraries are realizations of architecture in a specific hardware, networking, and software situation, which emphasize organization, acquisition, preservation, and utilization of information. According to *Ian Witten et al, digital library is* "A collection of digital objects, including text, video and audio, along with methods for access and retrieval, and for selection, organization and maintenance of the collection." DSpace is a digital library software by which we can build a Institution Repository (IR) including all of the above features with different languages. DSpace is developed by MIT and HP on March 2002 under the terms of the BSD open source license. DSpace, therefore, was designed as an open source application that institutions and organizations could run with relatively few resources. The intention to support interoperability (with DSpace implementers at other institutions, for example) led to the adoption of the Open Archives Initiative Protocol for Metadata Harvesting (OAIPMH); The OAI Registry includes DSpace, making its Dublin-Core- formatted metadata available to compatible harvesting code. In addition, DSpace chose to implement CNRI handles7 as the persistent identifier associated with each item to insure that the system will be able to locate and retrieve documents in the distant future. DSpace was also designed with a batch load submission feature to ease the loading of exiting collection and cut costs.



(Fig.1: A home page of DSpace digital Library)

3. UNICODE

Unicode provides a unique number for every character, no matter what the platform, no matter what the program, no matter what the language. A standard for representing characters as integers, unlike ASCII, which uses 7 bits for each character, Unicode uses 16 bits, which means that it can represent more than 65,000 unique characters. This is a bit of overkill for English and Western-European languages, but it is necessary for some other languages, such as Greek, Chinese and Japanese. Many analysts believe that as the software industry becomes increasingly global, Unicode will eventually supplant ASCII as the standard character coding format.

American Standard Code for Information Interchange(ASCII) is a seven-bit code proposed by American National Standards Institute (ANSI) in 1963 and approved in 1968. However, the present systems use a byte (8 bits) for a character, the 8th bit is used for the so-called extended ASCII. In other words, in 7 bits one can represent 128 values, whereas by using 8 bits, one can represent 256 values (i.e. 0-255). With the advent of Multimedia, much of the attention is paid to displaying or playing multimedia files. Many a format came into existence claiming better performance over its predecessors. However, when it comes to textual data, for a long time, there was no attempt to substitute ASCII for character representation. English has become a dominant language for communication of information and users were complacent with the 8-bit ASCII. Though an 8-bit approach to represent characters has been in use for scripts other than the Roman script, it became clear that one can not have a multi-lingual approach to represent various characters at a given time in a single screen or window. The reason is simple, in an 8-bit (1 byte), we can represent a maximum of 256 characters, whereas to represent characters belonging to various scripts require much more.

India belongs with several languages, cultures, customs and religions. The number of languages listed for India is 418. Of those, 407 are living languages and 11 are extinct. Eighteen are constitutionally recognized languages written in a variety of scripts. These are Hindi, Marathi, Gujarati, Punjabi, Bengali, Assamese, Manipuri, Nepali, Oriya, Telugu, Tamil, Malayalam, Kannada, Konkani, Sanskrit, Urdu, Kashmiri and Sindhi. We discuss the digital here Bengali language with Bengali script.

The Unicode Consortium was incorporated in January 1991, under the name Unicode, Inc., to promote the Unicode Standard as an international encoding system for information interchange, to aid in its implementation, and to maintain quality control over future revisions. The primary goal of the development effort for the Unicode Standard was to remedy two serious problems common to most multilingual computer programs. The first problem was the overloading of the font mechanism when encoding characters. Fonts have often been indiscriminately mapped to the same set of bytes. For example, the bytes 0x00 to 0xFF are often used for both characters and dingbats. The second major problem was the use of multiple, inconsistent character codes because of conflicting national and industry character standards. The Unicode Standard was designed to be:

- **Universal:** The repertoire must be large enough to encompass all characters that are likely to be used in general text interchange, including those in major international, national, and industry character sets.
- **Efficient:** Plain text is simple to parse: software does not have to maintain state or look for special escape sequences, and character synchronization from any point in a character stream is quick and unambiguous.
- **Uniform:** A fixed character code allows for efficient sorting, searching, display, and editing of text.
- **Unambiguous:** Any given 16-bit value always represents the same character.

The Unicode Standard defines three encoding forms that allow the same data to be transmitted in a byte, word or double word oriented format (i.e. in 8, 16 or 32-bits per code unit). All three encoding forms encode the *same* common character repertoire and can be efficiently transformed into one another without loss of data. The Unicode Consortium fully endorses the use of any of these encoding forms as a conformant way of implementingthe Unicode Standard.

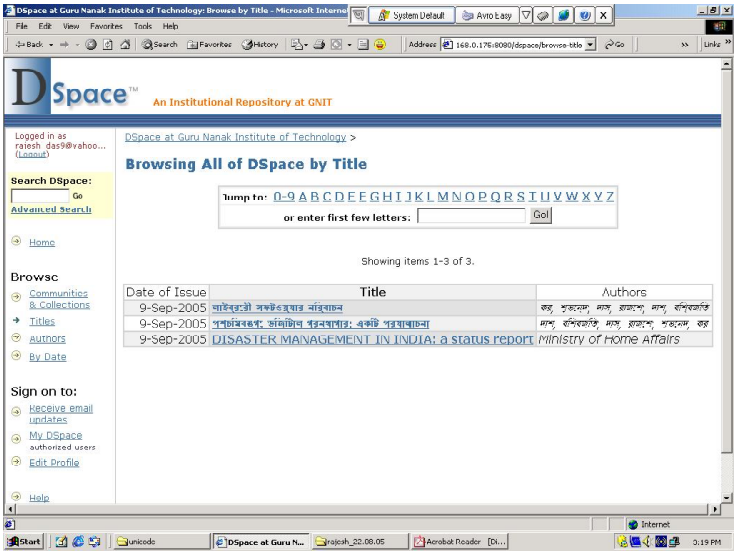
UTF-8 is popular for HTML and similar protocols. UTF-8 is a way of transforming all Unicode characters into a variable length encoding of bytes. It has the advantages that the Unicode characters corresponding to the familiar ASCII set have the same byte values as ASCII, and that Unicode characters transformed into UTF-8 can be used with much existing software without extensive software rewrites.

UTF-16 is popular in many environments that need to balance efficient access to characters with economical use of storage. It is reasonably compact and all the heavily used characters fit into a single 16-bit code unit, while all other characters are accessible via pairs of 16-bit code units.

UTF-32 is popular where memory space is no concern, but fixed width, single code unit access to characters is desired. Each Unicode character is encoded in a single 32-bit code unit when using UTF-32.

4. UNICODE AND DSPACE

To build a Unicode compatible document we need a Unicode supported word processor. Here we use word processor software for Bengali script “Avro Keyboard” which is developed by OmicronLab at Kolkata. It is Unicode compatible Bengali script software. In the Fig. 2 displays the Bengali language item out of three titles.

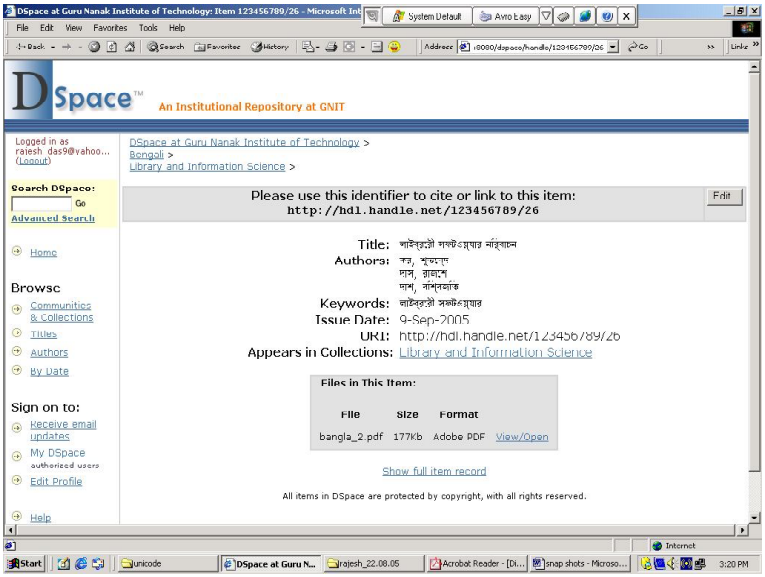


(Fig. 2: Display Bengali script titles of Bengali language document)

In the Fig. 3. It shows the authors' name in Bengali script

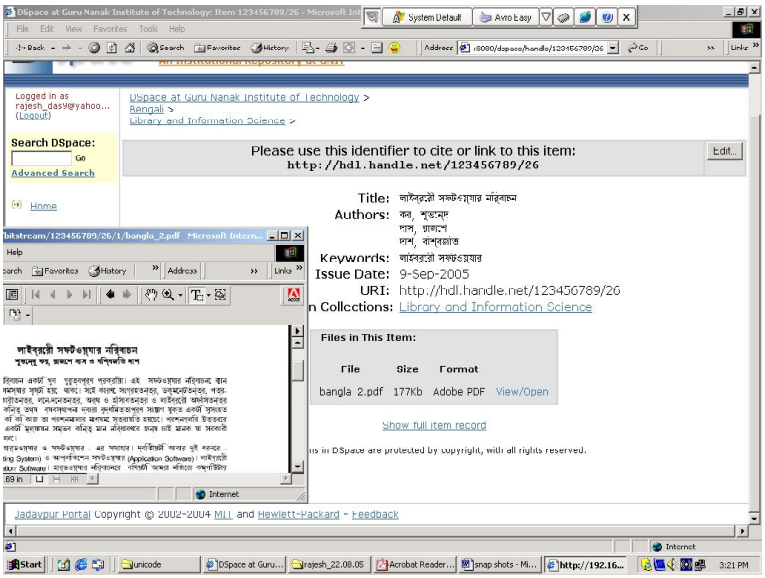
(Fig. 3: Display authors' name in Bengali script)

In Fig. 4 display the meta data in Bengali script. Display supports qualified Dublin Core meta data.



(Fig. 4: Display the meta data in Bengali script)

In the Fig. 5 the document is displayed in full text.



(Fig. 5: Display the document in full text)

5. CONCLUSION

At present, Library professionals have been handling information in various languages and when the information is to be digitized, Unicode becomes essential and handy.

It is imperative that operating systems, web browsers, word processors, database management systems and a host of application software should support Unicode for the realization of digital libraries.

It is a well recognized that digital libraries should be able to handle multimedia files. However, without multi-lingual capability, the digital libraries can only be partial. It becomes imperative that the content of the Internet and the digital libraries should be in Unicode as early as possible. Though Unicode is compatible with ASCII and one can develop software to convert ASCII files to UNICODE, the volume of information and data to be converted to Unicode becomes huge. Before we get into a problem similar to that of Y2K (if we are not already deep in that problem), the content of digital libraries and the Internet should adopt Unicode. India has long history and one of the oldest civilizations in the world with an awesome diversity of languages. 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 literature and scientific information over the centuries. The Internet and Unicode give a wonderful opportunity to various government and non-government organizations and to individuals to publish information in any of the Indian languages.

6. REFERENCES

1. LANGUAGES and scripts of India. from http://www.cs.colostate.edu/~malaiya/scripts.html search on 16/08/05
2. TEST for Unicode support in web browsers: Devanagari. from http://www.hclrss. demon.co.uk / unicode/devanagari.html search on 16/08/05
3. SALOMON, R.On the origin of the early Indian scripts: a review article.from http://www.ucl.ac.uk/ ~ucgadkw/position/salomon.html search on 16/08/05
4. THE Unicode standard, version 3.0. (2000). Massachusetts: Addison Wesley.

About Authors

Rajesh Das is a Student of PGDDLM of Department of Library and Information Science, Jadavpur University, Kolkata , West Bengal.

Biswajit Das is a Student of PGDDLM of Department of Library & Information Science, Jadavpur University, Kolkata , West Bengal.

Subhendu Kar is a Student of PGDDLM of Department of Library & Information Science, Jadavpur University, Kolkata , West Bengal.

Swarnali Chetterji is a Student of PGDDLM of Department of Library and Information Science, Jadavpur University, Kolkata , West Bengal.