International Journal of
Emerging Research in Management &Technology
ISSN: 2278-9359 (Volume-4, Issue-7)

Research Article

July 2015

Designing a Bengali Script based Retrieval System in Greenstone

[1]Debabrata Barman, [2]Anupratim Ghosh
[1]Junior Research Fellow, Department of Library and Information Science, University of Kalyani, Nadia, W.B., India
[2]Librarian, St. Augustine's Day School, Shyamnagar, W.B., India

Abstract-

This paper deals with the development of Bengali script based information retrieval system through the use of Free/Libre Open Source Software framework in general and Greenstone digital library software in particular. It briefly discusses characteristics of Indic scripts and their encoding in Unicode standard along with their encoding efforts in India. It uses Unicode as text encoding scheme, Avro as virtual keyboard for Bengali language, free open type fonts, rendering engine for Bengali script and applies these tools to design a multilingual Information Retrieval (IR) system in Greenstone DL software. The methodology is although meant for Bengal script can be applied to other Indic scripts for the purpose of designing multi-script/ multilingual information system.

Keywords: Digitization, Retrieval System, Bengali script based digital libraries

I. INTRODUCTION

With the development of digital libraries, objects can be accessed from users all over the world. Thus, digital libraries face the problem of establishing multilingual access to their collections. When researching multilinguality in information systems, three concepts are important to distinguish:

- Multilingual information access (MLIA),
- Multilingual information retrieval (MLIR) and
- Cross-language information retrieval (CLIR).

MLIA used as an umbrella term considers all aspects of multilinguality in information systems including accessibility, search, retrieval and inspection of objects regardless of the user or content language. Multilingual information retrieval describes systems that provide multilingual query functionalities and / or content more precisely, whereas cross-lingual information retrieval (CLIR) as part of information retrieval research focuses on the retrieval of documents in other languages than the query language. Up to date, most systems provide access to multilingual resources but only support monolingual search functionalities.

Dimensions of multilinguality in digital libraries can be classified according to three perspectives:

- User language,
- System language,
- Content language

The native or preferred languages of users as well as additional language skills influence user needs and requirements. The system language is represented by its interface. Multilingual systems provide localized interface representations for a selected set of languages. Besides the linguistic diversity of users and the respective interfaces, multilingual digital libraries also have to deal with content presented in several languages. The language of content in digital libraries can either be determined on the metadata or the object level. Especially for non-textual objects like images, only metadata information contains language information.

II. INDIC SCRIPTS

Indian is the world's second largest country in population, next only to China. It is a country with rich diversity in languages, customs and religions. India has 418 languages of which 407 are living and 11 are extinct. There are 18 constitutionally recognized languages written in a variety of scripts.

The Eighth Schedule of the Constitution of India, as of 1 December 2007, lists 22 languages, which are given in the table below together with the speaking population and the regions where they are used.

Table 1 Schedule of Indian Languages

Language	**Family**	**Speakers (in millions, 2001)**	**State(s)**
Assamese (Asamiya)	Indo-Aryan, North Eastern	13	Assam, Arunachal Pradesh
Bengali	Indo-Aryan, Eastern	83	West Bengal, Tripura, Assam, Andaman & Nicobar Islands, Jharkhand

Table 1 Schedule of Indian Languages

Language	**Family**	**Speakers (in millions, 2001)**	**State(s)**
Bodo	Tibeto-Burman	1.4	Assam
Dogri	Indo-Aryan, Northwestern	2.3	Jammu and Kashmir
Gujarati	Indo-Aryan, Western	46	Dadra and Nagar Haveli, Daman and Diu, Gujarat
Hindi	Indo-Aryan, Central	258–422	Andaman and Nicobar Islands, Bihar, Chhattisgarh, the National capital territory of Delhi, Himachal Pradesh, Jharkhand, Madhya Pradesh, Uttar Pradesh, Haryana and Uttarakhand
Kannada	Dravidian	55	Karnataka, Kerala, Tamil Nadu, Andhra Pradesh, Maharashtra
Kashmiri	Indo-Aryan, Dardic	5.5	Jammu and Kashmir
Konkani	Indo-Aryan, Southern	2.5	Goa, Maharashtra, Karnataka
Maithili	Indo-Aryan, Eastern	12.2	Bihar
Malayalam	Dravidian	33	Kerala, Lakshadweep, Puducherry
Manipuri(includes Meitei)	Tibeto-Burman	1.5	Manipur
Marathi	Indo-Aryan, Southern	72	Maharashtra, Goa, Dadra & Nagar Haveli, Daman and Diu
Nepali	Indo-Aryan, Northern	2.9	Sikkim, West Bengal
Odia	Indo-Aryan, Eastern	32	Odisha
Punjabi	Indo-Aryan, Northwestern	29	Chandigarh, Delhi, Haryana, Himachal Pradesh, Jammu, Punjab, Rajasthan, Uttarakhand
Sanskrit	Indo-Aryan	0.001	Uttarakhand
Santali	Munda	6.5	Santhal tribals of the Chota Nagpur Plateau (comprising the states of Bihar, Chhattisgarh, Jharkhand, Odisha)
Sindhi	Indo-Aryan, Northwestern	5	Sindh (now in Pakistan, Rajasthan, Kutch , Gujarat)
Tamil	Dravidian	61	Tamil Nadu, Andaman & Nicobar Islands, Kerala, Puducherry
Telugu	Dravidian	74	Andhra Pradesh, Telangana, Puducherry, Andaman & Nicobar Islands, Tamil Nadu
Urdu	Indo-Aryan, Central	52	Jammu and Kashmir, Telangana, Delhi, Bihar and Uttar Pradesh

(Sources: https://en.wikipedia.org/wiki/Languages_of_India)

III. ENCODING MULTILINGUALITY IR REQUIREMENTS IN UNICODE

The Unicode Standard is the universal character-encoding scheme for written characters and text. It defines a consistent way of encoding multilingual text that enables the exchange of text data internationally and creates the foundation for global software. The Unicode Consortium was incorporated in January 1991 to promote the Unicode standard as an international encoding system for information interchange. The Unicode Technical Committee (UTC) is the working group within the Consortium responsible for the creation, maintenance, and quality of the Unicode Standard. The UTC

follows an open process in developing the Unicode Standard and its other technical publications. During 1991, the Unicode Consortium and the International Organization for Standardization (ISO) recognized that a single, universal character code was highly desirable. A formal convergence of the two standards was negotiated, and they were merged into a single character encoding in January 1992 (Unicode Consortium, 2005). The present version i.e. Version 4.1 of the Unicode Standard is code-for-code identical to ISO/IEC 10646. This code-for-code identity holds true for all encoded characters in the two standards, including the East Asian (Han) ideographic characters. While modeled on the ASCII character set, the Unicode Standard goes far beyond ASCII's limited ability to encode only the upper- and lowercase letters A through Z. It provides the capacity to encode all characters used for the written languages of the world—more than 1 million characters can be encoded.

0.1. Multilingual Transliteration

Once being aware of the character codes, the approach towards transliteration can be taken. This feature is particularly advantageous with regard to Indian languages because all Indian languages use more or less the same consonants and vowels. Though some languages have some additional characters but in Content Representation in Devanagari UNICODE these are arranged in such a way that there is fixed difference in the value of same phonetic character belonging to different scripts. For example, the difference between all Telugu and Devanagari characters is 768.

0.2. Rendering Engine

Rendering is a complicated issue for Indian language processing. The complexity of rendering is easily noticeable from Fig. 1, where rendering of the Devanagari script is shown.

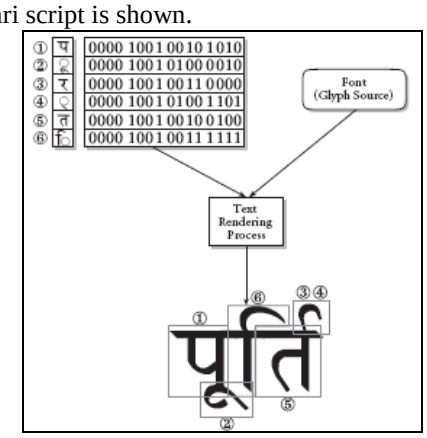

Fig. 1: Unicode Character Code to Rendered Glyphs
(Source – Unicode Standard Version 4.1)

0.3. Programming Environment

There are many additional features of Greenstone that lie outside the Librarian interface. Users can translate the interface into different natural languages. If they know HTML they can hook into Greenstone widgets like the full-text search mechanism or browsers from their own Web pages. If they know JavaScript they can incorporate browsing mechanisms such as image maps, and using Perl they can add entirely new browsing facilities, such as stroke-based or Pinyin-based browsing for Chinese. Some new requirements are best met by altering the Greenstone "receptionist" program, written in C++, to add new facilities at runtime.

0.4. Text Editor

A text editor is a computer program that lets a user enter, change, store, and usually print text (characters and numbers, each encoded by the computer and its input and output devices, arranged to have meaning to users or to other programs). Typically, a text editor provides an "empty" display screen with a fixed-line length and visible line numbers. You can then fill the lines in with text, line by line.

A popular text editor in IBM's large or mainframe computers is called XEDIT. In UNIX systems, the two most commonly used text editors are Emacs and vi . In personal computer systems, word processor is more common than text editors. However, there are variations of mainframe and UNIX text editors that are provided for use on personal computers. An example is KEDIT, which is basically XEDIT for Windows.

0.5. Database

Obviously, Unicode is not hardware or software. It allocates unique value to each character of the script and enables a single software product to process various characters from different scripts of the world. As a result, creation of multi-script databases requires not only Unicode-compliant operating system (OS) and other application programmes but also a Virtual Keyboard to enter multi-script records, Open Type Font (OTF) to support extended character sets and layout features, and Rendering Engine to display script specific conjuncts and ligatures properly.

In the beginning Unicode was a simple, fixed-width 16 bit encoding. Over the time, Unicode changed this fixed-width encoding style and presently allows three different forms of encoding to meet different requirements with databases.

- UTF-8 attempts to allow legacy systems to use Unicode by coding the characters in the ASCII character set with only eight bits, and encoding characters that are not in the ASCII character set with 16 bits. This is commonly used for Web pages.
- UTF-16 is supplementary characters outside the basic multilingual plane. It encodes most of the world's major languages in a fixed 16-bit character representation (2 bytes). This is the most common implementation.
- UTF-32 is an actually UCS 4, given a new name. It uses four bytes (32 bits) to encode all possible characters (rarely used).

IV. MULTILINGUAL RETRIEVAL SYSTEM IN BENGALI SCRIPT

This paper aims to develop an archive of Bengali script based documents and starts with designing an archive of Granthagar Patrika – a popular monthly journal of Library and Information Science in Bengal since 1953. First periodical on Library Science in Bengali was published by the Bengal Library Association in 1937 as 'Bengal Library Association Bulletin – Bangiya Granthagar Parishad Patrika'. It was a bilingual journal, which continued upto 1952. The Bengali journal "Granthagar" was first published in 1953 as quarterly and continued upto 1956. Since 1956 the Association publishes a monthly journal in Bengali with English Abstract named 'Granthagar'. Articles on Library and Information Science, memories of leading librarians, important circulars of the State and the Central Governments are published in the journal.

This paper describes that how to build a Bengali script based digital library using the Greenstone Digital Library (GSDL) Software – a comprehensive, open-source system for constructing, presenting, and maintaining digital collections. Collection is the articles of the 'Granthagar' which is published by the 'Bengal Library Association (BLA)' and can be built and rebuilt automatically in GSDL. The collections are easily maintainable and include effective full-text searching and metadata-based browsing facilities that are attractive and easy to use. Browsing utilizes hierarchical structures that are created automatically from metadata associated with the source documents. Collections can include text along with the title of the article, author, volume and the abstract using an easy to use tool called the Collector. Document in the collection is in Bengali language. Even the GSDL interface is available in many languages including Chinese and Arabic. The system is extensible and customizable i.e. software "plugins" can accommodate different documents and metadata types according to user requirement.

0.6. Use of Open Source Software

The software framework uses following open source software in developing the application environment:

Domain of Application	**Open Source Software**
Web server	Apache version 2.2.22 (http://httpd.apache.org/)
Programming environment	PERL (Version 5.14.2), PHP (Version 5.3.10) and Java Run time Environment (Version 1.7.0_80) (http://www.activestate.com; http://www.php.net; and http://java.sun.com/)
Application software (DL software)	Greenstone Digital Library Software (GSDL), Version 2.86 (http://greenstone.org/)
Virtual Keyboard (for Bengali script)	Avro Virtual Keyboard (http://www.omicronlab.com)
Rendering engine and Open type fonts (for Bengali script)	Pango & USP10.dll; Likhan, Bangla and Ekushey fonts (http://www.ekushey.org/)

All these software are open source software and may be downloaded freely from respective URLs. These are available against GPL (GNU Public License) and can be customized extensively as per the requirement of libraries.

0.7. Virtual Keyboard

Avro virtual keyboard an open source Windows based product of Omicron Lab, is selected for following reasons -

- The latest stable release of Avro (Version 1.7.7 release date 23/07/2014) is completely compatible with the latest Unicode version i.e. Version 4.1
- Preloaded with five keyboard layouts (Avro phonetic, Barnona, National, Unbijoy and Avro Easy) and allows customization of layout as per user requirements
- Facilitates a variety of typing options such as direct typing, typing from clip board and through mouse
- Supports phonetic typing on the basis of carefully crafted rules
- Allows custom configuration of the keyboard behaviour
- Users can turn on/off Bengali number typing from Numpad. This feature is extremely useful for Spreadsheet or Database management.

0.8. Programming Environment

Open source software in addition to multilingual tools like Avro Keyboard, rendering engine and open type fonts for Bengali language. These are Greenstone Digital Library (GSDL) software, PERL script based programming environment and Apache Web server. The system uses Dublin Core Application Profile for encoding language information.

0.9. Open Type Fonts

An open type font has two distinct advantages in a multilingual environment- its cross-platform compatibility and its ability to support widely expanded character sets and layout features. Open type fonts provide richer linguistic support and advanced typographic control such as glyph positioning, multiscript baselines, substitutional positioning, glyph classification and attachment, ligature definition and decomposition etc. A series of Bengali open type font s are produced by three Bengali language computational projects namely BengaliLinux of AnkurBangla Project (2003), Free Bangla Fonts Project (2004) or FBFP and Ekushey Project (2005). These are available for free downloading and includes Unicode compliant fonts like Akash, Likhan, Mitra, Rupali etc. (from AnkurBangla Project), Bangla, Ekushey Azad, Ekushey Durga, Ekushey Godhuli, Ekushey Mohua, Ekushey Sumit, Solaiman Lipi etc. from FBFP and Ekushey Project. All these fonts support Unicode 4.1 standard and USP10.DLL Uniscribe rendering engine, and thereby can be used in designing multilingual information system with necessary interface.

0.10. Text Editor

Using Text Editor (gedit) is the default GUI text editor in the Ubuntu operating system. It is UTF-8 compatible and supports most standard text editor features as well as many advanced features. These include Multilanguage spell checking, extensive support of syntax highlighting, and a large number of official and third party plugins. Gedit is suited for both basic and more advanced text editing and is released under the GNU General Public License.

0.11. Rendering Engine

Conjuncts and ligatures are the most font dependent of any scripts. They could be at different positions in different fonts. Uniscribe, which is called a rendering engine, should be using each font's glyph substitution tables to contextually render the characters.

- One Uniscribe engine, called USP10.DLL, is available as freeware to display properly open type Indic fonts on Windows operating system.
- And another Uniscribe engine use PANGO on linux. Pango is a text layout engine library which works with HarfBuzz shaping engine for displaying multi-language text. Full-function rendering of text and cross-platform support with APIs or 3rd party libraries, such as Uniscribe and FreeType, as text rendering backends.

0.12. Retrieval Engine:

MGPP (or MG++) is a reimplementation of MG, Managing Gigabytes, and is the default indexing tool used in the Greenstone. This tool provides search and retrieval of documents. There are two parts to using MGPP within Greenstone. One is collection building, and another is collection querying. Greenstone can build collections using MG or MGPP. The default is MG, but you can use MGPP by editing the collection configuration file(collect.cfg, found in the etc directory of a collection).

V. FEATURES OF BENGALI SCRIPTS BASE RETRIEVAL SYSTEM

In any Information Retrieval system searching is the most vital component for retrieval information. A search file is an essential component of a database which describes the document collection of an IR system. The formulation of the search strategy is the prerequisite for creation of a search file. There are several types of searching strategy in GSDL user interface,

0.13. Searching

There are two kind of searching strategy in GSDL.

0.13.1. Simple Search

The software allows search digital objects in simple search with GSDL user interface are:

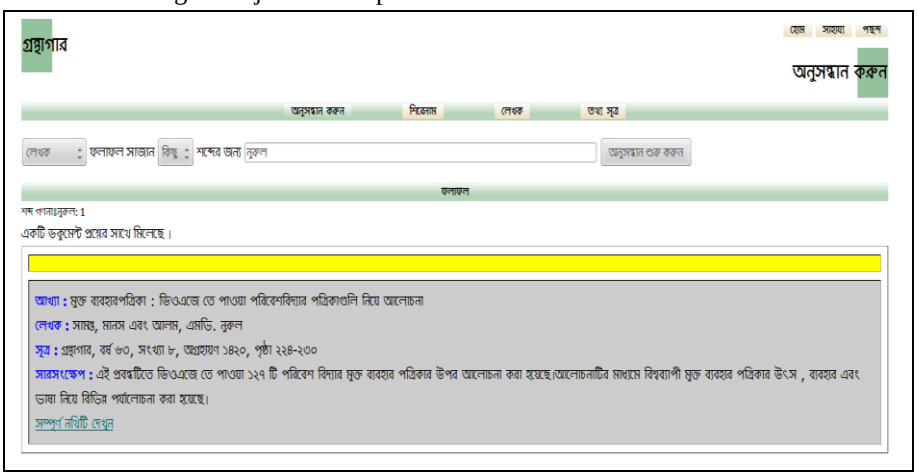

Fig. 2: Simple Search Technique of GSDL user interface

0.13.2. Advance Search

Advanced search is possible through the following options:

- Combined search through Boolean operators (AND, OR, NOT, XOR)
- Proximity operators, positional operators and relational operators
- Range search

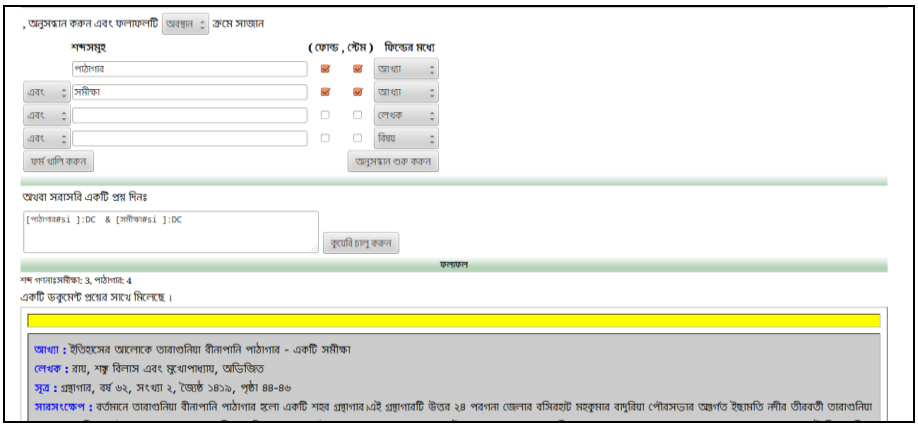

Fig. 3: Advance Search Technique of GSDL user interface

0.14. Browsing

Apart from these simple and advance search facilities this Unicode-compliant Bengali Script based Information Retrieval system also allows browsing of resources by name of title and author in alphabetically order. This facility is exhibited here as

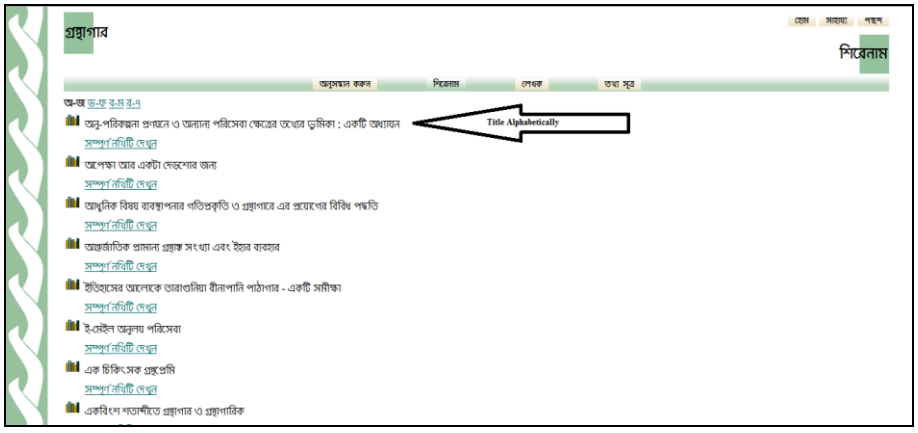

Fig. 4: Browsing Search Technique of GSDL user interface in Title Alphabetically

Fig. 5: Browsing Search Technique of GSDL user interface in Author Alphabetically

0.15. Search Operators Support

The discussion in foregoing paragraph shows application Boolean operators in two different search interfaces for Bengali script based documents in Greenstone. Apart from supporting field-level searching in Boolean query mode, this retrieval can support right truncation, left truncation, case-independent search, stem search. It also supports Fuzzy searching and application of Relational operators (>, <, >=, <=, =) and Positional operators (phrase, NEAR, ADJ etc.). The figure 3 shows that this Bengali script based retrieval system also supports weighted-term searching.

VI. CONCLUSION

This paper deals with the methodology for developing multilingual information retrieval system especially for Bengali language through the use of Unicode compliant software. The methodology profoundly depends on open source software. It uses Greenstone Digital Library Software (GSDL) as digital library management software, Apache web server, PERL programming environment, Avro virtual keyboard for Bengali, Open type fonts and necessary Rendering engine requires for target operating systems like Unices (Unix-like operating systems) or different Windows flavor like Windows XP, Windows 7, Windows 8 etc. All the abovementioned software and patches are downloadable freely from Internet along with source codes. This methodology represents processing and retrieval of Bengali language documents by the application of the said mechanism. This mechanism is primarily designed and developed for processing of Bengali language documents, but can be extended to manage all Indic scripts through the use of language specific virtual keyboards. Virtual keyboards for almost all the scheduled Indian languages are available freely from the web (see tdil.mit.gov.in). As this methodology largely depends on open source and freeware, Indian libraries of any type or size can adopt it to develop multilingual information retrieval system.

REFERENCE

[1] Acharya: Multilingual Computing for Literacy and Education. (2015). Unicode- A Brief Introduction. Retrived April 20, 2015. From http://www.acharya.gen.in:8080/multi_sys/unicode/uni.php?topic=introview
[2] AnkurBangla Project. (2003). BenglaLinux – RPM for open type Bengali fonts. Retrieved March 25, 2006, from http://www.sourceforge.net/projects/banglalinux /
[3] Apache Avro. (2015). Apache Avro™ Releases. Retrieved May 25, 2015, from https://avro.apache.org/releases.html
[4] Bureau of Indian Standards. (1991). Indian script code for information interchange IS 13194: 1991. New Delhi: Bureau of Indian Standard.
[5] Don, Katherine. (2015). MGPP: A search engine for XML documents User guide Retrieved June 18, 2015, from www.greenstone.org/manuals/mgpp_user.pdf
[6] Ekushey Project. (2005). Free open type Bengali fonts. Retrieved March 15, 2015, from http://www.ekushey.org/
[7] Free Bangla Font Project. (2004). Open type fonts for Bengali script. Retrieved March 15, 2015, from http://www.nongnu.org/freebangfont/
[8] Greenstone (software). (2015, May 21). In Wikipedia, The Free Encyclopedia. Retrieved June 15, 2015, from https://en.wikipedia.org/w/index.php?title=Greenstone_(software)&oldid=663426161
[9] Ibus-avro: Avro Phonetic Bangla typing for Linux Retrieved May 22, 2015, from http://linux.omicronlab.com/
[10] Khan, M.H. (2006). Avro version 3.1 – a free virtual keyboard for Bengali script compatible with Unicode 4.1 standard. . Retrieved May 25, 2015, from http://www.omicronlab.com/avrobangla/
[11] Languages of India. (2015, July 14). In Wikipedia, The Free Encyclopedia. Retrieved June 15, 2015, from https://en.wikipedia.org/w/index.php?title=Languages_of_India&oldid=671413668
[12] Mukhopadhyay, Parthasarathi (2006). Designing Web-enabled multilingual community information services: A FLOSS based framework for public libraries in West Bengal. Community information service – challenges and opportunities for libraries: Proceedings of the National Seminar of Department of Library and Information Science, Banaras Hindu University (March 20-22, 2006, Varanasi) (pp. 124-134). Varanasi: BHU.
[13] Mukhopadhyay, Parthasarathi. (2006). Public Library based Web-enabled Community Information System for Rural Development in India: Designing A FLOSS based Multilingual Prototype. Proceedings of the National Seminar on Open Source Movement – Asian Perspective, XXII, Roorkee, 2006. IASLIC, Kolkata. 2006. p. 251-258.
[14] Mukhopadhyay, Parthasarathi. (2006). Multi-script Information Retrieval System: A FLOSS based Prototype for Indic Scripts with Special Reference to Bengali Script. Proceedings of the Conference on Information Management in Digital Libraries. Indian Institute of Technology, Kharagpur. 2006. P. 305-316.
[15] Prasad, A.R.D. (2003). Creation of digital libraries in Unicode using Indian languages. In A R D Prasad (Ed.), Digital libraries: Theory and practice (pp. 105-114). Bangalore: Documentation Research and Training Centre.
[16] Technology Development for Indian Languages. (2015). About Indian languages. Retrieved June 18, 2015, from http://ildc.in/Bangla/Bindex.aspx
[17] Ubuntu: documentation (2015). Gedit. Retrived April 20, 2015, from https://help.ubuntu.com/community/gedit
[18] Unicode Consortium. (2005). The Unicode standard version 4.1. Massachusetts: Addison Wesley.
[19] Whatls.co. (2015). Text Editor. Retrieve April 20, 2015, from http://whatis.techtarget.com/definition/text-editor

