Designing Unicode-compliant Indic-script based Institutional Digital Repository with special reference to Bengali

Bijan Kumar Roy*, Subal Chandra Biswas**, Parthasarathi Mukhopadhyay***

ARTICLE INFO

Article history:
Received 15 May 2018
Revised 28 August 2018
Accepted 01 September 2018

Keywords:
Digital library,
Institutional Repository,
Multilingual Digital Library,
Multilingual Information Retrieval,
Indic-script

ABSTRACT

Local languages based information storage and retrieval system is essential for any online digital repository system. This paper reports the development of an interface in Bengali that allows users not only browsing and searching Indic-script based documents but also allows administrator performing various system level operations. This paper briefly describes the origin and key characteristics of Indic-scripts along with their encoding in Unicode standard with special reference to Bengali language. It also demonstrates the development processes of Bengal-script based information representation and retrieval (IRR) system viz. BURA (Burdwan University Research Archive) using different open standard and open source software (OSS) including different factors essential for building such successful Indic-script based multilingual digital libraries. The suggested strategies may help digital library developers to design an appropriate multi-script based information access services in any other Indic-script based languages.

1. Introduction

Language issues in any digital libraries are multifarious (Borgman, 1997). Local language based information storage and retrieval has become essential for the preservation of global heritage and culture. As a natural consequence of increasing globalization and the advent and growth of the Internet, digital libraries have been created that not only cross borders, but also languages. It deals with contents in more than one language and provides multilingual query access to a monolingual collection. The rapid growth of the non-English-speaking Internet population has created a need for better searching and browsing capabilities in languages other than English. However, existing search engines may not serve the needs of many non-English-speaking Internet users.

Almost 7097 languages are spoken in all over the world and English is the native language 118 countries having 378 millions speakers (Ethnologue, 2018). The Ethnologue is currently the

* Assistant Professor, Dept. of LIS, The University of Burdwan, WB (bkroy@lis.buruniv.ac.in) (Lead Author)
** Professor, Dept. of LIS, The University of Burdwan, WB (scbiswas_56@yahoo.co.in)
*** Professor, Dept. of LIS, University of Kalyani, WB (psmukhopadhyay@gmail.com) (Corresponding Author)
International Journal of Knowledge Content Development & Technology, 8(3): 53-67, 2018.
http://dx.doi.org/10.5865/IJKCT.2018.8.3.053

most comprehensive listing of the world's (mostly oral) languages. In this context, designing multilingual digital library is one solution and essential for Indian community. This language problem could be solved by creating a system of multilingual contents knowledge base system that might cover all Indian languages and will serve all regional community requirements. This paper presents the model viz. BURA (Burdwan University Research Archive) as a multilingual information representation and retrieval (MIRR) system and describes the methodology of designing such a Unicode-compliant Indic-script based institutional digital repository (IDR) system that supports processing and retrieving of non-English knowledge objects (here Bengali) in different languages through a multilingual user interface (here Bengali script-based interfaces).

The paper is not dealt with the over all large scale development of any search engine like Google (available in different Indian languages), Pipilika (www.pipilika.com), chorki (www.chorki.com) or Hindi khoj (http://hinkhoj.com/) etc. The objective of the paper is to develop Indic-script based user interface and retrieval mechanism in order to incorporate non-English knowledge objects generated by the Indian scholars working in different universities and research institutes.

2. Literature Review

World Wide Web (WWW) and Internet have linked all parts of the world and built a platform which we call '*digital earth*'. However, there are still barriers to overcome in order to benefit from those worldwide information resources. Displaying multilingual documents poses a big problem because of specific characteristics of each language and its character sets. Diekema (2012) reviewed the existing literature on multilingual digital libraries and provided an overview of this area. Another study (Wu, He, & Luo, 2012) surveyed academic users in order to identify their needs and expectations about multilingual information processing. The most important feature of any multilingual digital library system is that it allows browsing and searching across two or more different languages. This brings together collections from various countries, regions, cultures and provides access on a global scale (Yang, Wei, & Li, 2008; Maeda et al., 1998). In addition, it preserves cultural heritage and advance agriculture (Nichols et al., 2005).

Crossing the language barrier is one problem and is concerned with the translation of resources. Many authors have suggested developing multilingual information discovery system in order to make non-English content available to end users (Ghorab et al., 2011; Kaplan et al., 2014). Another study (Gibbon et al., 2004) reported the development of multilingual repository documenting in West African languages. Roy, Biswas and Mukhopadhyay (2017) developed Unicode-compliant Bengali script-based IDR system for Indian universities that supports integrated searching and browsing of Bengali language based resources. In another study, they proposed Bengali script-based interfaces that support integrated searching and browsing of Bengali language based resources with different search syntax (Roy, Biswas, & Mukhopadhyay, 2016).

Several other experts (Chung et al., 2004; Wang et al., 2006) presented a solution to the problem of missing dictionary terms in a query translation. The other challenges are data management and representation of information (Klavans & Schaüble, 1998); interoperability (Fox & Marchionini,

1998); development (Hutchinson et al., 2005); management and storage of content and metadata (Karvounarakis & Kapidakis, 2000). To solve this problem, Maeda et al. (1998) developed a technology to enable viewing of multilingual documents in a web browser. Another group of experts (Kramer, Nikolai & Habeck, 1997; McCulloch, Shiri, & Nicholson, 2005; Yang, Wei, & Li, 2008) put emphasis on achieving semantic interoperability to solve this problem.

Several studies have shared practical experience in implementing and managing multilingual digital library. Hutchinson et al. (2005) reported developing multilingual digital library '*International Children's Digital Library*' by the University of Maryland. Berkeley Public Library (www.berkeleypubliclibrary.org/multilingual_resources/) provided multilingual digital resources in eight languages, and had a multilingual catalogue search and multilingual reference service. Kopf et al. (2004) described another film archives project '*ECHO*' which holds documents in four languages, and it had cross-language search via a controlled vocabulary. Europeana (www.europeana.eu/portal/) which holds Europe's cultural heritage provides access to the materials in 27 different European languages.

3. Indic Scripts: Treatment in Unicode

India is rich in cultural and linguistic varieties. Indian community is based on several languages and dialect and Indian languages use a syllable as a basic linguistic unit. There are about 1650 dialects spoken by different communities and almost 10 Indic-scripts are in vogue (Vikas, 2001). Basically, scripts are of various types and its categorization depends on the way each character represents a phoneme (e.g. sound) or semantic unit (e.g. word, idea etc.). Indian languages owe their origin to Sanskrit and the Indic-scripts are all derived from ancient Brahmi script (4^{th} Century BC) and is said to have spawned more than 200 different scripts (Vikas, 2001). Though the origin of Brahmi script is not clear and the order of alphabets in all the scripts is similar. The Indic-scripts have a number of consonants, each of which represents a distinctive sound. Most of the major Indian languages (belongs to the Indo-Aryan and Dravidian group) use scripts which have evolved from the ancient Brahmi script (Sproat, 2002, 2003). At present, twenty-five languages and eighteen language scripts are constitutionally recognized such as Hindi, Marathi, Gujrati, Punjabi, Manipuri, Nepali and so on. Generally, Northern Indian languages belong to the family of Indo-European languages such as Hindi, Sanskrit. In the same way, South Indian languages belong to the Dravidian languages such as Tamil, Telegu (Roy, 2015). Fig. 1 & Fig. 2 is a visual representation of origin of Indian scripts along with some major Indian languages and related scripts.

It is found that Indian languages follow similar script and language grammars. One language may follow more than one script. For example, languages like Maithili, Punjabi, Nepali are written in multiple scripts. Even one script covers many languages. On the other hand, some languages use common script, especially Devanagari. The Devanagari script is used for writing classical Sanskrit and its modern historical derivative, Hindi. It (Devnagari) is the script used for several other languages such as Bhojpuri, Bihari, Hindi, Kashmiri, Konkani, Marathi etc. Similarly, Bengali script is used to represent languages of eastern India such as Bengali, Assamese, Vishnupriya Manipuri etc. Again, Brahmi script associated with many languages such as Bengali, Assamese, Manipuri and Sylheti.

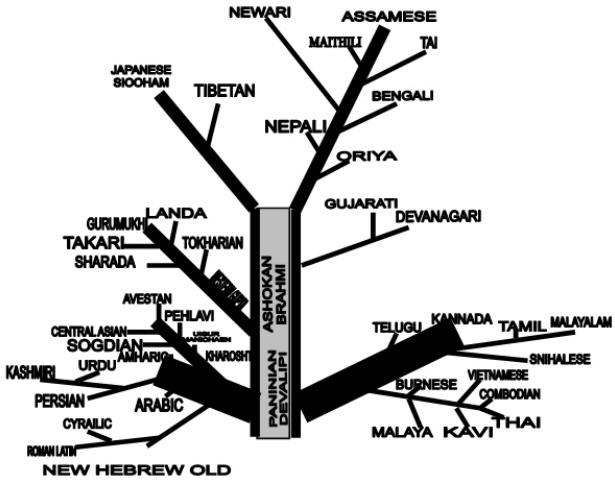

Fig. 1. Indic Scripts Origin
(Source: Vikas, 2005)

Sl. No.	Language	Script
1.	Hindi	Devanagari
2.	Sanskrit	Devanagari
3.	Marathi	Devanagari
4.	Konkani	Devanagari
5.	Nepali	Devanagari
6.	Maithili	Devanagari
7.	Sindhi	Devanagari
8.	Bodo	Devanagari
9.	Dogri	Devanagari
10.	Santhali	Devanagari, Ol Chiki
11.	Bengali	Bengali
12.	Assamese	Bengali
13.	Manipuri	Bengali, Meithei
14.	Gujarati	Gujarati
15.	Kannada	Kannada
16.	Malayalam	Malayalam
17.	Oriya	Oriya
18.	Punjabi	Gurmukhi
19.	Tamil	Tamil
20.	Telugu	Telugu
21.	Urdu	Perso-Arabic
22.	Kashmiri	Perso-Arabic

Fig. 2. Indian Languages and Scripts
(Source: Roy, 2014)

Unicode is basically a standard to represent universal character sets not the glyphs. For Indian languages, Unicode consortium adopted the 1988 version of ISCII - 8 (Indian Standard Code for Information Interchange) as the base for the 16-bit Unicode for allocating codes to different Indic-scripts. Here, code spaces for Indian-scripts are given below in Fig. 3.

Script	Assigned unique number
Arabic	U+0600 – U+06FF (01536 – 01791)
Devnagari	U+0900 – U+097F (02304 – 02431)
Bengali	U+0980 – U+09FF (02432 – 02559)
Gurumukhi	U+0A00 – U+0A7F (02560 – 02687)
Gujarati	U+0A80 – U+0AFF (02688 – 02815)
Oriya	U+0B00 – U+0B7F (02816 – 02943)
Tamil	U+0B80 – U+0BFF (02944 – 03071)
Telugu	U+0C00 – U+0C7F (03072 – 03199)
Kannada	U+0C80 – U+0CFF (03200 – 03327)
Malayalam	U+0D00 – U+0D7F (03328 – 03455)

Fig. 3. Codespace for Indian-scripts in Unicode
(Source: Unicode, version 11)

Recently, ISO (International Organization for Standardization) recommends renaming of “Bengali” script as “Bengali/Assamese” script as both the scripts share a large number of characters. In the new scripts some new set of symbols from the ‘Assamese script’ have been incorporated. Here, it is to be remembered that Assamese alphabets were in use even before the Bengali script but there was no separate slot for the Assamese script in the Unicode Standard. Now the chart would cover both the scripts.

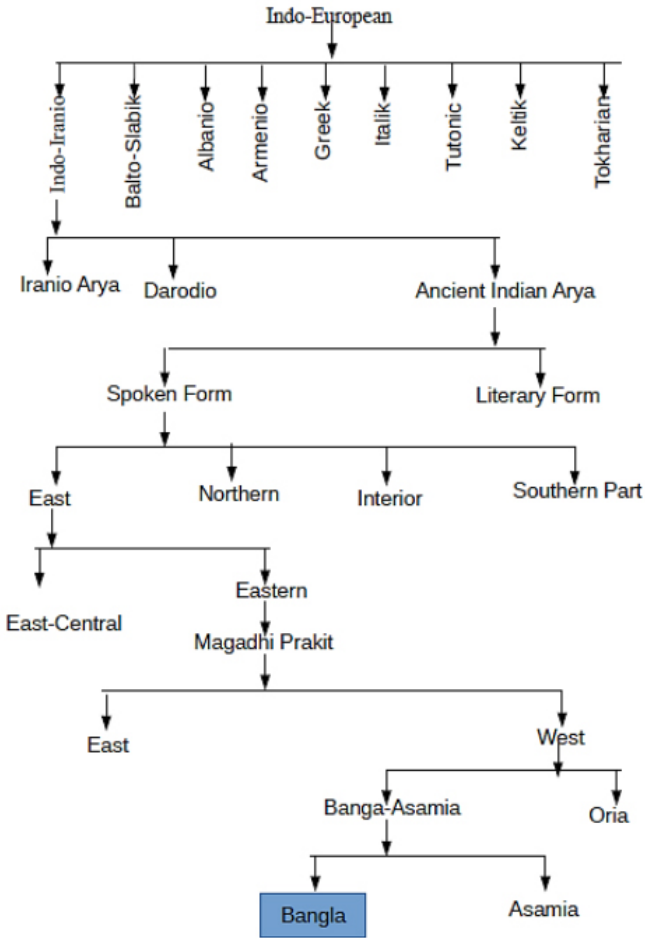

Fig. 4. Origin of Bengali Script

The above diagram (Fig. 4) shows the origin and development of Bengali Script (http://en.banglapedia.org/index.php?title=Bangla_Language). Fig. 5 gives the Unicode standard provided for Bengali Script. It is based on the 1988 version of ISCII (Indian Standard Code for Information Interchange) encoding (Unicode Consortium, 2016). The code space (e.g. Unicode Characters in the Bengali Block) for Bengali script (also known as Bangla) ranges from U+0980 to U+09FF. The Bengali script is a North Indian script closely related to Devanagari script and is used for writing Bengali and Assamese. Devnagari block runs from U+0900 to U+097F. It is based on Indian National Standard, ISCII. Like Devnagari block, Bengali follows the ISCII order which means that the corresponding characters in the two blocks are found at the corresponding position within their blocks. Though ISCII covers only ten Indian scripts and uses extended ASCII (American Standard Code for Information Interchange).

Fig. 5. Bengali-Script for Unicode (version 11.0)
(Source: Unicode)

4. Necessity of Indic-Script based Information Retrieval System

India is a multilingual country with millions of people speaking variety of languages. It has 418 languages of which 407 are living and 11 are extinct (Maitra, 2002). This diversity of languages is becoming barrier to understand and acquainted in digital world. As the country is diversified by languages, only 5% to 10% of population is aware of English language and can either read or write English. It was found in another report that less than 5% people can read and write English (Technology Development for Indian Languages Group, 2003). So, over 90% to 95% population is normally deprived of the benefits of English-based Information Technology (Vikas, 2005). As

a results, processing and retrieval of non-English knowledge resources have become a challenging task to the existing traditional keyword based textual information retrieval systems. It is due to the fact that earlier online information retrieval systems were based on ASCII (American Standard Code for Information Interchange) as text encoding standard and were unable to represent all the scripts of the world. Multiple languages could not be encoded by a single table. So, these systems were not able to process non-English language documents and language research in India was confined to the language translation only.

But situation started changing after the first Unicode project began to start in the late 1980s. Unicode is the first attempt to produce a standard for multilingual documents and made it possible to store and display hundreds of languages in their original script including many South Asian languages. Now-a-days, most of the major information retrieval systems are based on Unicode standard and are capable of handling all the scripts of the world (Daniels & Bright, 1998).

The Bengali language is currently the six most spoken language in the world (as on August, 2018) with roughly 243 million speakers (https://www.statista.com/statistics/266808/the-most-spoken-languages-worldwide/) (Fig. 6). As an Indian language, Bengali ranks 2nd position after Hindi where 8.03% people speak in Bengali (http://censusindia.gov.in/2011Census/C-16_25062018_NEW.pdf).

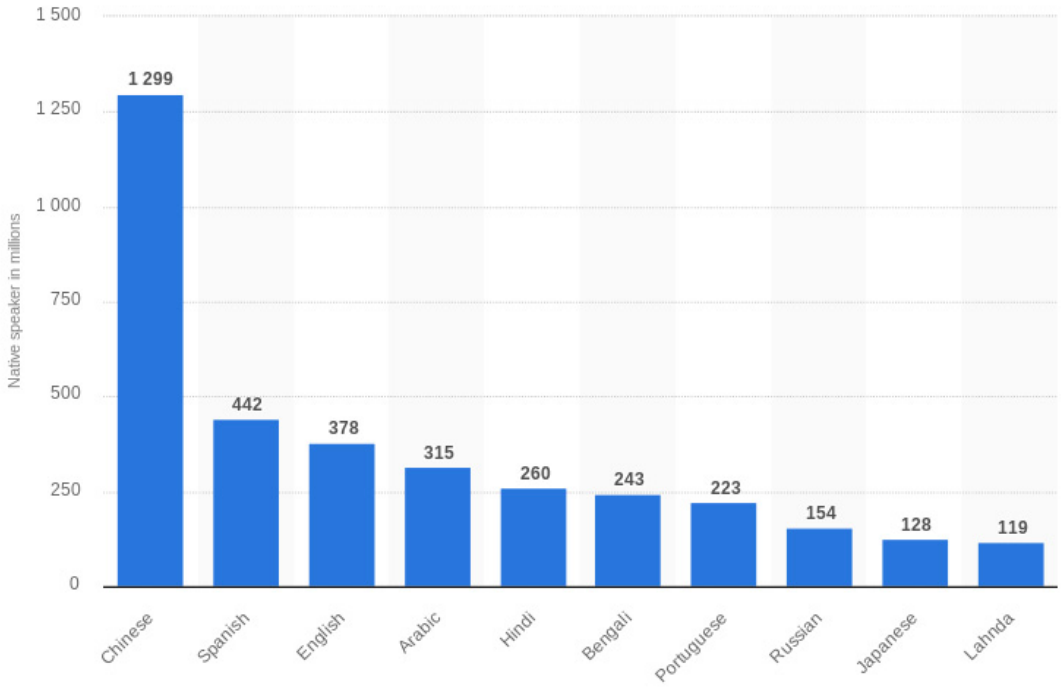

Fig. 6. Most spoken languages worldwide

Due to this language diversity, an Indic-script based information retrieval and representation (IRR) system is essential and may be a solution to the academic community in processing and retrieving non-English knowledge objects. This system is expected to process all types of public funded research

outputs (e.g. dissertations, theses, reports) produced by academicians, scientists, researchers in different regional languages. Otherwise, these open knowledge resources other than English languages will remain inaccessible even to the stakeholders due to absence of appropriate retrieval mechanisms.

5. Methodology

This section gives a brief theoretical overview of the methodology followed for the development of the model viz. BURA. A wide range of criteria have been considered before designing this kind of Indic-script based multilingual digital archive. The model has emerged on the basis of a number of initiatives, projects etc at national and international level. This section addresses the design of a Bengali-script based information representation and retrieval (IRR) system viz. BURA using different open source software (OSS) and open standard technologies in different layers and levels of its implementations. For this purpose, DSpace (http://www.dspace.org/) software has been used and configured to have interfaces in Bengali language. It also uses Linux (Ubuntu) as operating system, Apache as web server, PostgreSQL as relational database, Java as programming/scripting language. All these software were integrated and deployed in DSpace to make it Unicode compatible. This part mainly focuses on different techniques and technologies used in designing BURA software framework and also explain various steps involved to configure DSpace to handle non-English content and creation of user interface in Bengali. The development process of this Bengali-script based framework involves the following six major steps such as - a) selection and installation of software used in different layers; b) development of Basic Cluster (through Linux-Apache-PostgreSQL-Java); c) selection and installation of repository software (here DSpace); d) development of multilingual cluster; e) translation of messages in Bengali; and f) performing and configuring other repository related tasks.

All these steps may be categorized in three broad groups viz. *Database/Retrieval related steps* (deals with selection of default schema 'UTF8' as native character set), *Servlet engine related* (URI encoding in '*server.xml*' file) and *Interface related steps* (concerned with translation of messages in '*message.properties*' file). The whole process may schematically be illustrated as below (Fig. 7).

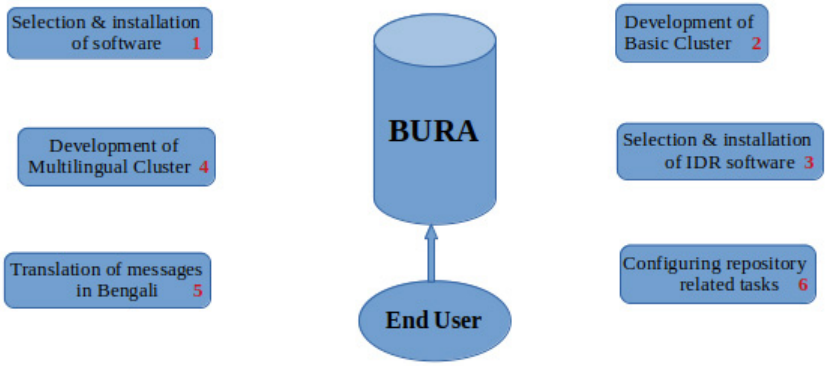

Fig. 7. Development of the Model

STEP 1

The first logical step of designing Indic-script based user interface in Bengali is making UTF-8 as default character encoding scheme or native character set for DSpace database in PostgreSQL (Fig. 8). It uses UTF-8 encoding internally. This step maintains a link between the software framework and the database used by the system (here DSpace).

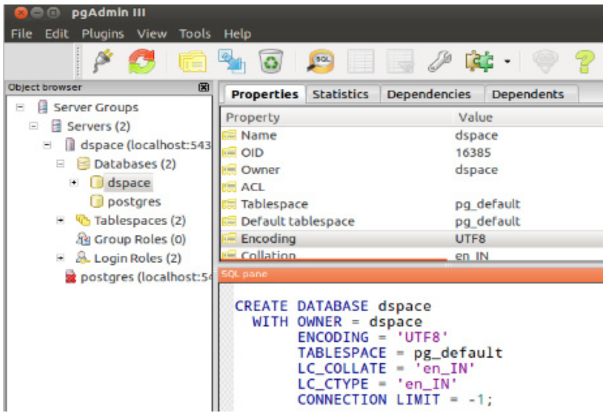

Fig. 8. DSpace database with UTF-8 as native character set

STEP 2

Making necessary changes and translations of messages (from English to Bengali) in *Messeage. properties* file (available in the location - /webapps/dspace/WEB-INF/classes/Messages.properties) is the next logical step to make the default interface (in English) in Bengali script. By default, for the English interface the extension of the file is 'en'. Fig. 9 shows the English language messages for *Submission Aspect* section of DSpace.


```
                 Submission Aspect
      !!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!-->

<!-- general submission aspect messages -->
<message key="xmlui.Submission.general.mydspace_home">MyDSpace Home</message>
<message key="xmlui.Submission.general.go_mydspace">Go to MyDSpace Home</message>
<message key="xmlui.Submission.general.submission.title">Item submission</message>
<message key="xmlui.Submission.general.submission.trail">Item submission</message>
<message key="xmlui.Submission.general.submission.head">Item submission</message>
<message key="xmlui.Submission.general.submission.previous">&lt; Previous</message>
<message key="xmlui.Submission.general.submission.save">Save &amp; Exit</message>
<message key="xmlui.Submission.general.submission.next">Next &gt;</message>
<message key="xmlui.Submission.general.submission.complete">Complete submission</message>
<message key="xmlui.Submission.general.workflow.title">Item submission</message>
<message key="xmlui.Submission.general.workflow.trail">Item submission</message>
<message key="xmlui.Submission.general.workflow.head">Item submission</message>
<message key="xmlui.Submission.general.showfull">Show full item record</message>
<message key="xmlui.Submission.general.showsimple">Show simple item record</message>
<message key="xmlui.Submission.general.default.title">Submission</message>
<message key="xmlui.Submission.general.default.trail">Submission</message>

<!-- org.dspace.app.xmlui.submission.CollectiovViewer -->
<message key="xmlui.Submission.CollectionViewer.link1">Submit a new item to this collection</message>

<!-- org.dspace.app.xmlui.submission.Navigation -->
<message key="xmlui.Submission.Navigation.submissions">Submissions</message>
```

Fig. 9. Original Message File in English

This default message property file has been translated in Bengali by using carefully translated message headings through the use of Unicode-compliant text editor. The following figure (Fig. 10) shows the Bengali equivalents of English language messages for *Submission Aspect* section of DSpace.

```
                    Submission Aspect

        !!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!-->

<!-- general submission aspect messages -->
<message key="xmlui.Submission.general.mydspace_home">আমার ডিস্পেস মূলপাতা</message>
<message key="xmlui.Submission.general.go_mydspace">আমার ডিস্পেস মূলপাতা যান</message>
<message key="xmlui.Submission.general.submission.title">তালিকাভুক্ত বিষয়ের দাখিল</message>
<message key="xmlui.Submission.general.submission.trail">তালিকাভুক্ত বিষয় দাখিল করুন</message>
<message key="xmlui.Submission.general.submission.head">তালিকাভুক্ত বিষয় দাখিল করুন</message>
<message key="xmlui.Submission.general.submission.previous">পূর্ববর্তী</message>
<message key="xmlui.Submission.general.submission.save">সংরক্ষণ করুন; প্রস্থান করুন</message>
<message key="xmlui.Submission.general.submission.next">পরবর্তী; </message>
<message key="xmlui.Submission.general.submission.complete">দাখিল করা সম্পূর্ণ হয়েছে</message>
<message key="xmlui.Submission.general.workflow.title">তালিকাভুক্ত বিষয়ের দাখিল</message>
<message key="xmlui.Submission.general.workflow.trail">তালিকাভুক্ত বিষয় দাখিল করুন</message>
<message key="xmlui.Submission.general.workflow.head">তালিকাভুক্ত বিষয় দাখিল করুন</message>
<message key="xmlui.Submission.general.showfull">সমগ্র বিষয়বস্তু দেখুন</message>
<message key="xmlui.Submission.general.showsimple">নির্দিষ্ট সাধারণ তালিকাভুক্ত বিষয় দেখানো হবে</message>
<message key="xmlui.Submission.general.default.title">দাখিল করুন</message>
<message key="xmlui.Submission.general.default.trail">দাখিল করুন</message>

<!-- org.dspace.app.xmlui.submission.CollectiovViewer -->
<message key="xmlui.Submission.CollectionViewer.link1">এই সংগ্রহে একটি নতুন তালিকাভুক্ত বিষয় দাখিল করুন</message>

<!-- org.dspace.app.xmlui.submission.Navigation -->
<message key="xmlui.Submission.Navigation.submissions">দাখিল করুন</message>
```

Fig. 10. Modified Message File in Bengali

STEP 3

Making necessary changes in *server.xml* file (Configuration file of Tomcat) is required to ensure UTF-8 enabled transactions between Tomcat server and DSpace database. The first figure (Fig. 11) shows default *server.xml* file and the second figure (Fig. 12) shows the modified *server.xml* to support Unicode based database transactions.

```
<!-- Define a non-SSL HTTP/1.1 Connector on port 8080 -->

<Connector
port="8080"    maxHttpHeaderSize="8192"
          maxThreads="150" minSpareThreads="25" maxSpareThreads="75"
          enableLookups="false" redirectPort="8443" acceptCount="100"
          connectionTimeout="20000" disableUploadTimeout="true" />

<!-- Note : To disable connection timeouts, set connectionTimeout value
   to 0 -->
```

Fig. 11. Programme box: Original entry in server.xml file

```
<Connector
port="8080"          maxHttpHeaderSize="8192"
        maxThreads="150" minSpareThreads="25" maxSpareThreads="75"
        enableLookups="false" redirectPort="8443" acceptCount="100"
        connectionTimeout="20000" disableUploadTimeout="true"
 URIEncoding="UTF-8"/>
```

Fig. 12. Programme box: Modified entry in server.xml file

Here, step1 and step 2 are related with user interface and retrieval and are concerned with browsing and searching of resources. Step 3 is concerned with multilingualism e.g. making postgreSQL database Unicode compliant.

6. User Interface and Retrieval

This Indic-script based software framework is based on open standard and open source software (OSS) and architecture is basically a combination of Linux-Apache-PostgreSQL and Java. Like other existing multilingual information representation and retrieval (MIRR) system, this system has the following key features - i) *supports universal character set or in other words helps in contents development and contents access in Unicode-compliant Bengali script*; ii) *allows integrated browsing (by different search syntax) and searching (including advanced searching) multilingual environment and dissemination of multiple media such as text, audio, video objects etc*; iii) *allows easy access and submission of objects of different forms and formats*; iv) *filtering of search results by language and category*; v) *switching the user interface language (here from English to Bengali and Bengali to English by selecting the corresponding language tab) at any time during the interaction*; vi) *can be integrated with any network at national level*; and vii) *providing multilingual access to content as well as allowing metadata description*.

After combing above three categories of work, it produces the following retrieval facilities such as browsing, searching, metadata and text selection etc. This model supports browsing resources by different search syntax such as by author, by title, by subject etc. Fig. 13 is the example of browsing resources by 'title'.

Advanced search interface (Fig. 14) supports searching through Boolean operators (e.g. AND, OR, NOT) and also allows for metadata selection.

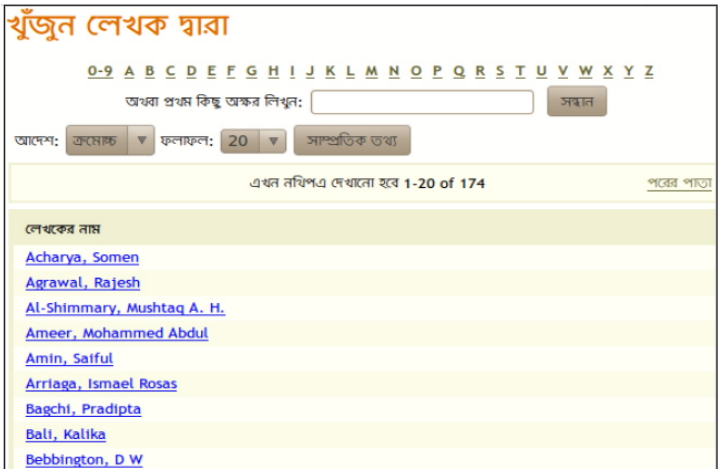

Fig. 13. Browsing resources by author

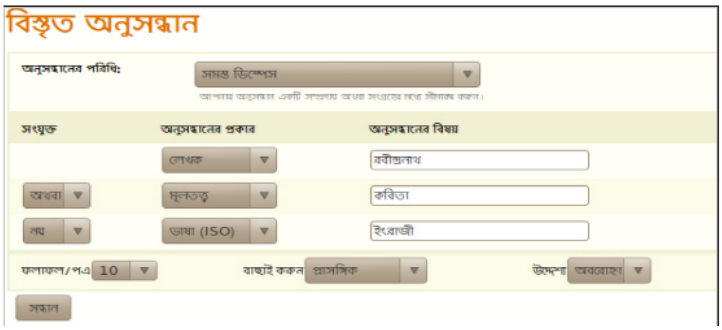

Fig. 14. Advanced search interface

7. Concluding Remarks

Due to the enormous growth of non-English informational objects, the demand for MIRR System is growing and Library and Information Science (LIS) professionals could play a great role in designing Indic-script based information retrieval system using Unicode compliant Free/Libre Open Source Software (FLOSS). There are possibly no work has been done in Indian languages specially in Bengali and a very few works are available for foreign languages. It has been proved that people like to have information in original language and if the services are provided in local languages (e.g. browsing and searching), it is an added value service on behalf of the library. But challenges are diverse such as localization and language processing, lack of fonts and character codes, technical matter like software design, interoperability and there are further issues to explore. This is basically an experiment with Bengali-script but this mechanism may be extended in any other Indic-script based languages such as Hindi, Telugu, Tamil etc. The required thing is that only the heavy work related with the transliteration of messages but URI encoding and PostgreSQL work tool will remain

the same. Only language specific user interface is required to be developed. This mechanism may be used in Bangladesh and may cover other 24 Indian languages in designing IDR for their university with some modification in *message.properties* file but other steps will remain the same. So, the development of Web-enabled Unicode-compliant system should be a mandatory parameter for any online digital IRR system in India which will support and promote researchers working in different languages by providing seamless access to non-English documents.

References

Borgman, C. L. (1997). Multi-Media, Multi-Cultural, And Multi-Lingual Digital Libraries. *D-Lib, 3*(6). Retrieved from http://www.dlib.org/dlib/june97/06borgman.html

Chung, W., Zhang, Y., Huang, Z., Wang, G., Ong, T. H., & Chen, H. (2004). Internet searching and browsing in a multilingual world: An experiment on the Chinese Business Intelligence Portal (CBizPort). *Journal of the American society for information science and technology, 55*(9), 818-831.

Daniels, P., & Bright, W. (1998). The world writing systems. Massachusetts: Addison Wesley.

Diekema, A. R. (2012). Multilinguality in the digital library: A review. *The Electronic Library, 30*(2), 165-181.

Ethnologue (2018). Languages of the World. Retrieved from https://www.ethnologue.com/statistics/size

Fox, E. A., & Marchionini, G. (1998). Toward a Worldwide Digital Library. *Communications of the ACM, 41*(4), 29-32.

Ghorab, M. R., Leveling, J., Lawless, S., O'Connor, A., Zhou, D., Jones, G. J., & Wade, V. (2011, September). Multilingual adaptive search for digital libraries. In *International Conference on Theory and Practice of Digital Libraries* (pp. 244-251). Springer, Berlin, Heidelberg.

Gibbon, D., Ahoua, F., Gbéry, E., Urua, E. A., & Ekpenyong, M. (2004). WALA: A Multilingual Resource Repository for West African Languages. In *LREC*.

Hutchinson, H. B., Rose, A., Bederson, B. B., Weeks, A. C., & Druin, A. (2005). The international children's digital library: a case study in designing for a multilingual, multicultural, multigenerational audience. *Information Technology and Libraries, 24*(1), 4-12.

Kaplan, A., Sándor, Á., Severiens, T., & Vorndran, A. (2014). Finding quality: a multilingual search engine for educational research. In *Assessing Quality in European Educational Research* (pp. 22-30). Springer VS, Wiesbaden.

Karvounarakis, G., & Kapidakis, S. (2000). Submission and repository management of digital libraries, using WWW. *Computer Networks, 34*(6), 861-872.

Klavans, J. L., & Schauble, P. (1998). NSF-EU Multilingual Information Access: imagine information storage, access, and presentation without first translating into the user's native language. *Communications of the ACM, 41*(4), 69-70.

Kopf, S., Haenselmann, T., Farin, D., & Effelsberg, W. (2004, June). Automatic generation of video summaries for historical films. In *Multimedia and Expo, 2004. ICME'04. 2004 IEEE*

International Conference on (Vol. 3, pp. 2067-2070). IEEE.

Kramer, R., Nikolai, R., & Habeck, C. (1997). Thesaurus federations: loosely integrated thesauri for document retrieval in networks based on Internet technologies. *International Journal on Digital Libraries, 1*(2), 122-131.

Maeda, A., Dartois, M., Fujita, T., Sakaguchi, T., Sugimoto, S., & Tabata, K. (1998). Viewing multilingual documents on your local web browser. *Communications of the ACM, 41*(4), 64-65.

Maitra, D. (2002). Languages and scripts of India. Retrieved from http://www.cs.colostate.edu/~maitra/scripts.html

McCulloch, E., Shiri, A., & Nicholson, D. (2005). Challenges and issues in terminology mapping: a digital library perspective. *Electronic Library, The, 23*(6), 671-677.

Nichols, D. M., Witten, I. H., Keegan, T. T., Bainbridge, D., & Dewsnip, M. (2005). Digital libraries and minority languages. *New Review of Hypermedia and Multimedia, 11*(2), 139-155.

Roy, B. K. (2015). *Institutional Digital Repositories: From Policy to Practice*. LAP LAMBERT Academic Publishing.

Roy, B. K., Biswas, S. C., & Mukhopadhyay, P. (2016). Open access repositories for Indian universities: towards a multilingual framework. *IASLIC Bulletin, 61*(4), 150-161.

Roy, B. K., Biswas, S. C., & Mukhopadhyay, P. (2017). BURA: an open access multilingual information retrieval and representation system for Indian higher education and research institutions.

Sproat, R. (2002). Brahmi scripts. In Constraints on Spelling Changes: Fifth Inter-national Workshop on Writing Systems, Nijmegen, The Netherlands.

Sproat, R. (2003, December). A formal computational analysis of indic scripts. In *International symposium on indic scripts: past and future, Tokyo*.

Technology Development for Indian Languages Group (2003). About Indian languages. Retrieved from http://tdil.mit.gov.in/home.asp

Unicode Consortium (2016). The Unicode standard version 5.1. Massachusetts: Addison Wesley.

Vikas, O. (2005, May). Multilingualism for cultural diversity and universal access in cyberspace: an Asian perspective. In *Thematic meeting for the world summit on the information society* (pp. 1-49).

Wang, J. H. et al. (2006). Exploiting the Web as the multilingual corpus for unknown query translation. *Journal of the American Society for Information Science & Technology, 57*(5), 660-670.

Wu, D., He, D., & Luo, B. (2012). Multilingual needs and expectations in digital libraries: A survey of academic users with different languages. *The Electronic Library, 30*(2), 182-196.

Yang, C. C., Wei, C. P., & Li, K. W. (2008). Cross-lingual thesaurus for multilingual knowledge management. *Decision Support Systems, 45*(3), 596-605.

Yang, C. C., Wei, C. P., & Li, K. W. (2008). Cross-lingual thesaurus for multilingual knowledge management. *Decision Support Systems, 45*(3), 596-605.

[About the authors]

Bijan Kumar Roy, M.Com, MLIS, PhD is Assistant Professor in Library and Information Science, The University of Burdwan, West Bengal, India. He started his career as full time research fellow and later joined as Librarian in Government-aided College in 2009. His research interest includes open access, open source software, digital repository.

Subal Chandra Biswas, b. 1955, M.A. (Economics), MLIS, Ph.D. (Loughborough) has recently retired as Professor of LIS, The University of Burdwan, West Bengal, India. Recipient of Commonwealth Scholarship (UK), 1985-1989. He has an experience in teaching and research of more than three decades both at home and abroad. Research interests include information seeking, information retrieval, and public libraries. Has supervised more than a dozen doctoral theses.

Parthasarathi Mukhopadhy, MLIS, PhD is Professor in Library and Information Science, University of Kalyani, Kalyani, West Bengal, India. His research interest includes open access resource organization, open source applications in library organization and multilingual information retrieval. He is presently associated with two mega digital library projects in India namely National Digital Library Initiative and National Virtual Library Initiative.