

Retrieval in Texts with Traditional Mongolian Script Realizing Unicoded Traditional Mongolian Digital Library

Garmaabazar Khaltarkhuu and Akira Maeda
Graduate School of Science and Engineering, Ritsumeikan University
1-1-1, Noji Higashi, Kusatsu, 525-8577 Shiga, Japan
garmaabazar@gmail.com and amaeda@is.ritsumei.ac.jp

Abstract

This paper discusses our approaches to create a digital library on traditional Mongolian script using Unicode. Also we introduce system architecture of a digital library that stores books and materials of historical importance written in traditional Mongolian which contain history of 1,000 years and are important part of Mongolian culture. Specifically, we propose a technique that will allow users to search traditional Mongolian unicoded texts with keywords in modern Mongolian Cyrillic characters. To accomplish our goal, we used Greenstone digital library system and it is based on a unicoded traditional Mongolian script. We created a traditional Mongolian digital library with Golden History (Altan Tobci in Mongolian) - chronological book of ancient Mongolian Kings and their history. We approved our system's effectiveness by experiment.

Keywords: Traditional Mongolian Script, Digital library, Unicode, Information Retrieval, Cyrillic input.

1. Introduction

The main purpose of this research is to develop a technique to keep over 1,000 years old historical records written in traditional Mongolian script including history of Chinggis Khaan for futures use, to digitize all existing records and to make those valuable data available for public viewing and screening.

There are over 50,000 registered manuscripts and historical records written in traditional Mongolian script stored in the National Library of Mongolia [1]. About 21,100 of them are handwritten documents. There are many more manuscripts and books in traditional Mongolian script stored in libraries of other countries such as China, Russia and Germany. Despite the importance of keeping 1,000 years old historical materials in good conditions, the Mongolian environments for material storage is not satisfactory to keep historical records for a long period of time. We believe that the most efficient and effective way to keep and protect old historical materials while making it publicly available is to digitalize historical records and create a digital library. There are several obstacles in the traditional Mongolian script information processing for instance IME is not available; differs from the modern Mongolian language and other spoken-Mongolian dialects; letters have at least three different variations which are decided by their position in a word or depend on the preceding letter with which it forms a ligature. Thus this paper introduces some techniques to build Mongolia-specific digital

library for documents written in the traditional Mongolian script. Especially we propose retrieval technique of the traditional Mongolian text using modern Mongolian query. We also describe conversion technique of traditional Mongolian script in Unicode basic character to presentation character. In addition we integrate our techniques to Greenstone Digital Library.

2. Traditional Mongolian script and Mongolian language

As mentioned before the main purpose of this paper is to explore opportunities to build a traditional Mongolian script digital library. One of the biggest problems is that the traditional Mongolian script differs from the modern Mongolian language. At present, people use dictionaries between the traditional Mongolian written words and modern Mongolian words. Mongolia introduced a new writing system (Cyrillic) in 1946. This has been a radical change and alienated the traditional Mongolian language.

2.1 Unicode and feature of traditional Mongolian script

The traditional Mongolian script character code set has been placed in Unicode at the range of 1800-18AF [2]. It is not enough to solve problems in processing information in Mongolian. Problems described below still exist. The traditional Mongolian writing system is known to be quite different from the western systems as well as the CJK (Chinese, Japanese and Korean) writing systems.

- Traditional Mongolian script is written vertically from top to bottom in columns advancing from left to right. This directional pattern is unique among existing scripts. Thus, general operating systems fail to correctly display traditional Mongolian script. Despite much effort, it remains a major problem. At present HyperText Markup Language (HTML) and Cascading Style Sheets (CSS) support vertical text displaying from right to left in web browsing, but do not support vertical text displaying from left to right, as it is the case for traditional Mongolian script.
- Traditional Mongolian characters are written in succession, meaning that depending on where a letter is placed in a word, it may have different forms. There are at least three different forms for each letter and some letters have a dozen different forms. Those are called: isolate, initial, middle, and final form. All of these forms are decided by their position in a word. Some sample is shown in Figure 1.

Mongolian letter (transliteration)	Isolate	Initial	Medial	Final
ᠠ (A)	ᠠ	ᠠ	ᠠ ᠠ ᠠ	ᠠ
ᠡ (OE)	ᠡ	ᠡ	ᠡ ᠡ ᠡ	ᠡ
ᠢ (L)		ᠢ	ᠢ	ᠢ

Figure 1. Initial, medial and final forms of Mongolian script letters [3].

The form of a letter may also depend on the preceding letter with which it forms a ligature (shown in Figure 2). Each letter has a basic form as well as some possible variations of forms, while certain combinations of letters combined form ligatures.

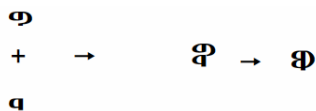


Figure 2. Mongolian script ligatures [3].

The Unicode standard includes only the basic character sets, special punctuation symbols and numerals, but does not explicitly encode the variant forms or the ligatures, although the correct variant form or ligature can, in most cases, be determined from the context. We will explain later about principles how to generate variant and how to display traditional Mongolian text correctly.

3. Related work

Man et al. introduced a method for electronizing the traditional Mongolian script and its application to text retrieval [4]. They developed a transcript of traditional Mongolian and Roman characters and used Roman characters input for Mongolian text. Their research was not based on Unicode and merely used Roman characters transcripts and stored Roman-character-based content. To display Roman characters text as traditional Mongolian text, they used JavaScript. Search function was working without any problem since stored contents are Roman characters.

4. Traditional Mongolian Script Digital Library (TMSDL) Realization

4.1 Overview

In this section we will introduce traditional Mongolian script digital library with Cyrillic interface [5]. We utilized Greenstone Digital Library (GSDL), developed by New Zealand Digital Library (NZDL) Consortium at the University of Waikato to build TMSDL and created sample collection. GSDL is a suite of open source software for building and distributing digital library collections. GSDL uses Unicode and XML-compliant format internally, and supports indexing of large collection of information including multimedia. The basic structure of our system is shown in Figure 3.

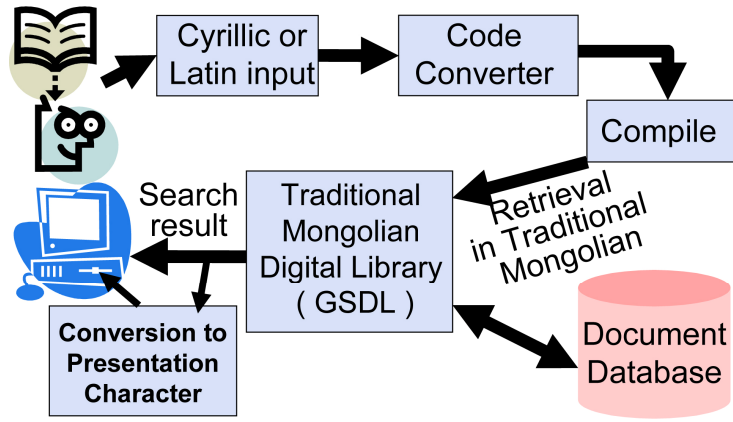


Figure 3. Traditional Mongolian script digital library system. General architecture consist of user search input interface, converter, compiler, dictionary, GSDL core and display interface.

4.2 Components

4.2.1 Cyrillic and Latin code converter and Traditional Mongolian retrieval in GSDL

One of the main functions of a digital library is the search engine. Input Method Editor (IME) is not available for traditional Mongolian script text input. On the other hand, text input in Cyrillic is available. If we take into account that in Mongolian it is relatively easy to find Cyrillic and Latin IME, these scripts should be used in our digital library's search engine. User will input keyword(s) in Cyrillic or in Latin alphabet. Since we have chosen GSDL as the base system, the user interface has to be web-based.

Content is stored in unicoded traditional Mongolian basic forms and not variations, since Unicode standard includes only the basic character set, but does not explicitly encode the variant forms or ligatures, while the correct variant form or ligature can be determined from the context. Thus our system converts modern Mongolian query to traditional Mongolian and displays correct variant form, when user input Cyrillic search text. Built collection is shown in the Figure 5. When a query in modern Mongolian is inputted, converter function converts the modern Mongolian text to a traditional Mongolian query. Next, the traditional Mongolian query is retrieved from the GSDL.

Our approach is not to touch GSDL source code and do not want to modify the standard macro files [6]. Instead we created collection specific macro file extra.dm (/collect/<collname>/macros) to add or override our functions. This is done by the below code sequences (shown in Figure 4).

```

package query _queryform_ { <form name=QueryForm method=get
  action="_gwcgi_" onSubmit="toconvert();">...
  _dummpagescriptextra_ { function toconvert()
    \{ ...our conversion algorithm ... \} }
}

```

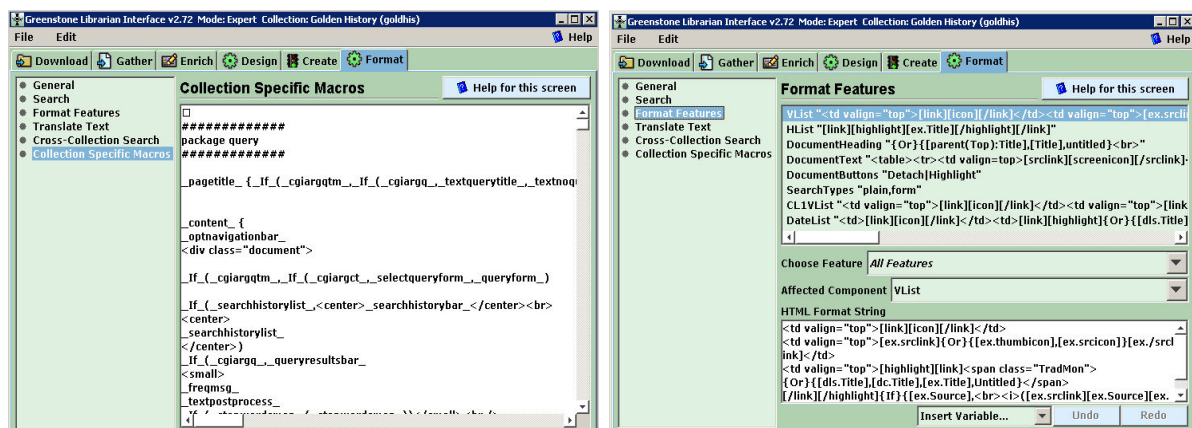


Figure 4. Collection Specific Macro Integration in GUI Administration page of the GSDL



Figure 5. Traditional Mongolian Collection on the GSDL: Golden History of Mongolian Kings, Cyrillic and Latin code converter and Traditional Mongolian retrieval in Cyrillic input

4.2.2 Code converter with Unicode in the display interface

If we store letter's variant forms, indexing and searching functions will become complicated [7]. Therefore we use code converter to display already stored basic characters correctly. Example of conversion is shown in Figure 6. There are control-symbols encoded that can be used to resolve ambiguities in few cases where the context rules are inadequate. These control-symbols can also be used to override the default forms if it is required. The control-symbols are the Mongolian free variant selectors (180B-^{FV}_{IS1}, 180C-^{FV}_{IS2}, 180D-^{FV}_{IS3}) for selecting alternative variants of a given positional form, and the Mongolian vowel separator 180E -^M_{VS1}. The Mongolian vowel separator serves as a distinguisher of the vowels "A" and "E". It is because, these two vowels are written exactly the same when they are placed at the

end of a word. Examples shown in Figure 7 illustrate the use of the Mongolian vowel separator^[MVS].

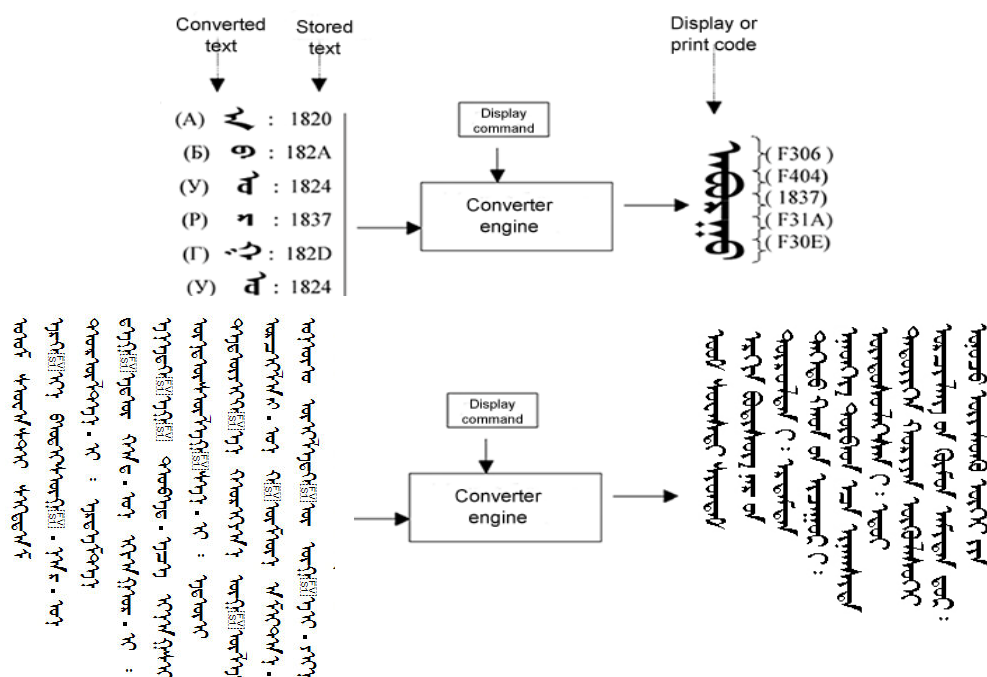


Figure 6. Converter engine for traditional Mongolian script

Character sequence	Display	Character sequence	Display
... ? [MVS] ᠠᠨ	ᠠᠨ	ᠠᠨ	ᠠᠨ
... ᠠᠨ [MVS] ᠠᠨ	ᠠᠨ	ᠠᠨ	ᠠᠨ
... ᠠᠨ [MVS] ᠠᠨ	ᠠᠨ	ᠠᠨ	ᠠᠨ
... ᠠᠨ [MVS] ᠠᠨ	ᠠᠨ	ᠠᠨ	ᠠᠨ
... ᠠᠨ [MVS] ᠠᠨ	ᠠᠨ	ᠠᠨ	ᠠᠨ

Figure 7. Examples illustrate the use of the Mongolian vowel separator

The Mongolian free variant selectors are used to distinguish different variants of the same positional form of a character. They modify only the character immediately preceding them and will have no effect on the character following. Examples shown in Figure 8 illustrate the use of the Mongolian free variant selectors.

Character sequence	Example of use	Character sequence	Example of use
ᠠᠨ [FV S1]	ᠠᠨ	ᠠᠨ	ᠠᠨ
... ᠠᠨ [FV S1]	ᠠᠨ	... ᠠᠨ	ᠠᠨ
ᠠᠨ [FV S1] ...	ᠠᠨ (traditional form)	ᠠᠨ ...	ᠠᠨ
... ᠠᠨ [FV S1]	ᠠᠨ	... ᠠᠨ	ᠠᠨ

Figure 8. Examples illustrate some uses of the free variant selectors

We developed an algorithm to display the traditional Mongolian characters correctly using control

characters and/or basic characters. This is one of the most important parts of traditional Mongolian script digital library. Then we integrated our algorithm which is written in JavaScript into GSDL without touching GSDL source code and modifying the standard macro files.






4.3 Experiment result

After creating our collection in the Greenstone Digital Library System, we run several experiments to test our system effectiveness by inputting Cyrillic query. Then we compared our retrieval result with statistics of linguistic analysis book of source document The Golden History (Altan Tobci) [8].

First, we tested with major grammars of traditional Mongolian to check whether our converter works correctly or not. Then we selected several noun and numeral from most repeatedly used words and run our experiment. There are 127 words that repeated more than 20 in the Golden History.

In our experiment, retrieval was successful for given traditional Mongolian words and retrieved word counts were matched with source document. Thus our algorithm and modern Mongolian to a traditional Mongolian query converter works perfectly for commonly-used traditional Mongolian words. Some sample search keywords are shown in Table 1.

Table 1: Example of query result

Cyrillic Input	Traditional mongolian query (meaning)	Retrieved	All numbers[8]
Эзэн	 (lord)	146	146
Жил	 (year)	86	86
Энэ	 (this)	86	86
Зарлиг	 (order)	65	65
Хан	 (prince)	61	61

5. Conclusion and future work

In this paper we introduced system architecture for a traditional Mongolian script digital library. We proposed possible methods for traditional Mongolian text displaying and converting user query that will enable digital library system to search traditional Mongolian text with keywords in modern Mongolian characters. Those are main parts of traditional Mongolian script digital library. We successfully demonstrated our collection in traditional Mongolian digital library. The Golden History (year 1604, 164pp) - Chronological book of ancient Mongolian Kings and their history is now available in the traditional Mongolian digital library. Retrieval was successful for commonly-used traditional Mongolian words. Our future work will focus on irregular words and

grammar. Words that have different meanings but are written and pronounced exactly the same in modern Mongolian can have different forms when converted to traditional Mongolian. Consequently, word sense disambiguation must be considered.

Lastly, although traditional Mongolian script character codes are already available in Unicode, recently we have realized some researchers are lack of using it. On the other hand, some researches incorrectly used traditional Mongolian script's control-symbols. Control-symbols, variant shapes and ligatures are the most important understanding of traditional Mongolian script. Usage of those symbols is causing additional problems in Mongolian information processing. Thus promoting Unicode and utilizing traditional Mongolian script in Unicode is essential.

References

1. Тунгалаг, Д.: Монгол улсын үндэсний номын сан дахь монголын түүхийн гар бичмэлийн номзүйн судалгаа, 1-р боть. Тайм принтинг, Улаанбаатар (2005) (in Mongolian)
2. The Unicode Consortium: The Unicode Standard 4.0. Addison-Wesley, Boston San Francisco New York (2003)
3. Erdenechimeg, M., Moore, R., M., Namsrai, Yu.: UNU/IIST Technical Report No. 170 – Traditional Mongolian Script in the ISO/Unicode Standards (1999)
4. Man, D., Fujii, A., Ishikawa, T.: A Method for Electronizing the Traditional Mongolian Script and Its Application to Text Retrieval. The IEICE Transactions D-II Vol. J88-D-II No.10 (2005), 2102-2111(in Japanese)
5. Garmaabazar, Kh., Maeda, A.: Retrieval Technique with the Modern Mongolian Query on Traditional Mongolian Text, In Proceedings of the 9th International Conference on Asian Digital Libraries (ICADL2006), (2007), 478-481
6. Garmaabazar, Kh., Maeda, A.: Building a Digital Library of Traditional Mongolian Historical Documents, In Proceedings of the 7th ACM/IEEE Joint Conference on Digital Libraries(JCDL 2007), (2007), 483
7. Насан-Урт, С.: Монгол хэл бичгийн сураг занги боловсруулах онол практикийн зарим асуудал. Улаанбаатар (2004) (in Mongolian)
8. Choimaa, Sh., Shagdarsuren, Ts.: “Qad-un undusun quriyangyui altan tobci”, (Textological Study) (2002) (in Mongolian)