



## An Evaluation of Multi-language/multi-script Functions in KOHA

**Naicheng Chang**

General Education Center

Tatung University

Taipei, China

[ncchang@ttu.edu.tw](mailto:ncchang@ttu.edu.tw)

**Yuchin Tsai**

National Center for High-performance Computing

National Applied Research Laboratories

Taipei, China

[Thomas@nchc.org.tw](mailto:Thomas@nchc.org.tw)

**Meeting:** 135. UNIMARC

---

WORLD LIBRARY AND INFORMATION CONGRESS: 75TH IFLA GENERAL CONFERENCE AND COUNCIL

23-27 August 2009, Milan, Italy

<http://www.ifla.org/annual-conference/ifla75/index.htm>

---

### **Abstract:**

*As far as systems for library management are concerned, a recent trend has been to develop Open Source Software (OSS). Library automation software uses a number of standards which are almost universally applied across the library world. One of these, UNIMARC, has various dialects. This paper looks at implementing UNIMARC and Chinese MARC (CMARC) on Koha, one of the longer-established OSS packages. It will determine to what extent the various features of UNIMARC and CMARC which are not present in MARC21 such as linking fields can be supported in Koha and what is required to implement them if they are not available.*

### **1. Background**

The Integrated Library System (ILS) environment has changed considerably over the last few years. A recent trend has been to develop Open Source Software (OSS) ILSs. Libraries turn to OSS solutions mainly because several OSS library management systems such as Koha have

been considered modern and mature systems that would fulfill libraries' needs; OSS are more open to customization to meet the special demands of libraries; OSS features emerge from the user community that have contracted or developed and contributed them so that other libraries can use and benefit from them. Furthermore, libraries using OSS have more support options than those using proprietary software (Breeding, 2007).

Library automation requires a higher level of text processing than many information retrieval systems because it needs to take into account characters with diacritics. Computers began in the United States where there is predominantly used the basic character set of latin characters without diacritics. However, one needs to know if eleve will retrieve élève or not. One needs to be able to distinguish between eleve and élève. The situation becomes even more complex with non-latin alphabets. Unicode has been developed for this purpose.

Library automation software uses a number of standards which are almost universally applied across the library world so every library systems should incorporate them. One of these standards, arguably the most important as it allows the sharing of records created in one library around the rest of the world, is MARC (Machine-Readable Cataloging). Unfortunately, it has various dialects, one of which is used extensively around the world. It is called UNIMARC, Universal MARC Format. In Taiwan, a sub-dialect of this, Chinese MARC (CMARC), is the most widely used machine-readable format among libraries. The most common dialect of all is that developed originally by the US Library of Congress as LC MARC which is now known as MARC21. All OSS packages are likely to implement this format, but other dialects have more advanced features since they have been developed more recently than LC MARC/MARC21. For example, UNIMARC implements more subfields so has a higher level of granularity than MARC21, and has different methodologies of linking which need implementing separately by any software.

UNIMARC in its 3<sup>rd</sup> edition (2008) has been updated recently to take into account new changes in the internationally-accepted record structure to take into account Unicode. In 2004, the National Central Library (Taiwan) hosted an unofficial Unicode Workgroup with the purpose of seeking solution of issues of multiple internal codes, that is, Chinese Character Code for Information Interchange (CCCII) and Big5, the most widely used Chinese internal codes among libraries in Taiwan. (Mao and Hsu, 2006)

Koha was originally developed for use with MARC21 which is used around the English-speaking world. In the following section, we will discuss the implementation of UNIMARC and CMARC on Koha to evaluate to what extent the various features of UNIMARC and CMARC which are not present in MARC21 such as linking fields (e.g. the

methodologies for linking from a record of a serial to its later or earlier titles) can be supported in Koha, and what is required to implement them if they are not available.

The Koha system was originally developed for the Horowhenua Library Trust by Katipo Communications in 1999. Koha went into use at the Nelsonville Library (Athens County, Ohio) in the fall of 2003. According to Breeding (2008), Koha ranks as the first full-featured open source ILS and serves the most number of libraries, mainly public libraries, in the US. Koha is only at the early stage both in the UK and in Taiwan, yet the future development in the two countries could be promising.

## 2. A Koha UNIMARC/CMARC/MARC21 Testbed

The testbed we are working with is to take Koha as our library system to examine the various features in MARCs; that is UNIMARC, CMARC and MARC21. In the testbed, 24 ISO2709-compliant UNIMARC bibliographic records, and 100 each of CMARC and MARC21 bibliographic records were tested.

### Defining Koha Framework

Before putting Koha to work, we defined the MARC fields and subfields framework using tools provided by the Koha system. Presently, Koha officially provides default MARC21 and UNIMARC templates. It is possible to redefine the cataloguing framework according to the practices of the library. The Koha-Taiwan team has developed a free CMARC template which is currently used in most Koha libraries in Taiwan (Mao, 2006). A redefining example in CMARC below in Figure 1 shows on the left-hand the default Koha CMARC, and in the right-hand, a subfield \$s has been redefined in Field 700, and so is displayed in the catalogue record. The data in subfield \$s remains in Koha even without being redefined.

**MARC biblio : 11 (黃金時代)**  
With Framework : Default

0 1 2 6 7 8

700 #1 - 人名 -- 主要著者  
a 標目主體 司馬  
b 副標目 光  
4 著作方式 原著

702 #1 - 人名 -- 其他著者  
a 標目主體 柏楊  
4 著作方式 譯

**MARC biblio : 11 (黃金時代)**  
With Framework : Default

0 1 2 6 7 8

700 #1 - 人名 -- 主要著者  
s 朝代 宋  
a 標目主體 司馬  
b 副標目 光  
4 著作方式 原著

702 #1 - 人名 -- 其他著者  
a 標目主體 柏楊  
4 著作方式 譯

Figure 1: Redefining Koha Framework in CMARC

In the testbed, we defined the testbed to be based on UNIMARC 3<sup>rd</sup> edition (2008), CMARC 4<sup>th</sup> edition with update 2001 and MARC21 1999 edition with update 2008.

## Importing MARC Records

We encountered no difficulty in importing bibliographic records in the three MARC formats except that the default encoding system in Koha is UNICODE ( that is UTF-8 here), and in Taiwan, the most accepted internal codes are CCCII and Big5. Special programming is required here in order to import the Chinese bibliographic records into Koha and display properly the data (Tsai, 2007). An example of UNIMARC importing data is shown in Figure 2. Clicking the far right end Bib numbers, the system will show the MARC records.

### Manage Staged MARC Records › Batch 3

<b>File name</b>	nchc_0513_chi.iso
<b>Comments</b>	import marc from nchc(none)
<b>Staged</b>	2009-03-17 15:21:53
<b>Status</b>	staged
<b>Matching rule applied</b>	No matching rule in effect
<b>Action if matching record found</b>	create_new
<b>Action if no match found</b>	create_new
<b>Item processing</b>	ignore

  

<b>New matching rule</b>	Do not look for matching records ▼
<b>Action if matching record found</b>	Add incoming record ▼
<b>Action if no match found</b>	Add incoming record ▼
<b>Item processing</b>	Ignore items ▼

Apply different matching rule

Import into catalog

Page 1 [2](#) [3](#) [4](#)

#	Citation	Status	Match?	Bib
1	<a href="#">Statistical theory of heat : Brenig, Wilhelm. (3540510362 )</a>	imported	no_match	<a href="#">1</a>
2	<a href="#">Computational methods for kinetic models of magnetically confined plasmas / (0387134018 )</a>	imported	no_match	<a href="#">2</a>
3	<a href="#">Linear kinetic theory and particle transport in stochastic mixtures / Pomraning, G. C.</a>	imported	no_match	<a href="#">3</a>
4	<a href="#">Physics of space plasmas : Parks, George K. (0201508214 )</a>	imported	no_match	<a href="#">4</a>
5	<a href="#">A computational approach to chemistry / Hirst, David M. (0632024313   0632027436 )</a>	imported	no_match	<a href="#">5</a>
6	<a href="#">Topological methods in chemistry / Merrifield, Richard E. (0471838179 )</a>	imported	no_match	<a href="#">6</a>
7	<a href="#">Algorithms for chemists / Zupan, Jure. (0471921734 )</a>	imported	no_match	<a href="#">7</a>
8	<a href="#">Factor analysis in chemistry / Malinowski, Edmund R. (0471530093 )</a>	imported	no_match	<a href="#">8</a>
9	<a href="#">Theoretical and computational models for organic chemistry / (0792313143 )</a>	imported	no_match	<a href="#">9</a>
10	<a href="#">Mathematical frontiers in computational chemical physics / (0387967826 )</a>	imported	no_match	<a href="#">10</a>

Figure 2: UNIMARC Records

## MARC View and Simple View

Bibliographic data can be displayed in MARC format, in simplified form, or in ISBD format, in both the librarian interface and the OPAC. Before viewing, we defined display formats of fields and subfields. In Figure 3, a UNIMARC record displays three options of viewing: Normal View, MARC View and ISBD View.



**Figure 3: ISBD View of a UNIMARC Record**

## Language Encoding and Transmission

It is essential to have knowledge of the encoding system before working on the Koha system. ISO2709 defines the length of fields, yet the length varies due to the characters. For example, Chinese characters take 2 bytes which is different from Western characters, and this causes extra work, that is special programming, when working on Koha. This situation is commonly seen in Asian countries, such as China, Japan and Korea.

In the testbed, we tested three MARC formats and proved that libraries need to define the MARC framework before importing their data, and the result showed that there is no difficulty for any MARC formats to work on Koha properly.

## 3. Comparison and Evaluation

There are many subtle differences between the formats but they are all very closely related to each other because they all act as a carrier for Anglo-American Cataloguing Rules. UNIMARC was developed with a view to moving away from the catalogue card concept inherent in MARC originally; the tags of MARC are generally ordered as in a catalogue card with main heading, description, subject access and other added entries (tracings) in that order. UNIMARC has a more logical tag hierarchy ordered by function. After the coded identifiers/standard numbers and coded data elements (MARC21 begins with codes and follows with identifiers), we find description (ISBD), notes (as in ISBD), linking (which

probably follows notes because the traditional output of a link is a note) subject, and then at 700 name access points. Individual tags within the blocks are different. However, these distinctions are in many senses trivial. More important as far as compatibility is concerned are issues relating to granularity at the subfield level and the methods of linking. As far as implementation by software is concerned, the two features which need to be tested for individual implementation are linking techniques and coded data fields. Linking is a complex activity. Coded data fields are intricate as far as data entry is concerned. That is because the fields require careful counting of codes to enter and to avoid this being onerous we need a tailor-made entry methodology. Coded fields are important for multi-lingual multi-script functions. This data is best stored as coded data. Work has been done on defining language in webpages with a view to helping with retrieval through search engines and it is just as important to define language of text in a catalogue. Additionally, end users are usually interested in the language of a bibliographic item and even of the original work. These data are stored in coded form. If a system cannot deal with a script, there needs to be transliteration with appropriate information to enable conversion in a system where it is available and to permit information retrieval.

## **Comparison**

### **A. Coded Data Fields**

We tested Coded Data Fields, that is, fixed-length fields, in UNIMARC and CMARC on field 100, 101, 105, 110, 115, 116, 117, 120, 121, 125, 126, 128, 129, 130,135,140,141, and in MARC21 on field 007 and 008. We discovered that in the basic Koha system only are found default Coded Data Fields 100, 101 and 105 with subfield \$a in UNIMARC and CMARC as showed in Table 1, yet full subfields in 007 and 008 in MARC21 shown in Figure 4 by clicking the icons in the far right end. This implies that systems librarians need to enable the subfields by redefining the Koha system framework when working with UNIMARC or CMARC; whilst all the subfields of Coded Data Fields are available there on MARC21. For technical support terms of view, this is an advantage of MARC21 when implementing Koha into libraries because not much extra work is needed.

<b>tagfield</b>	<b>tagsubfield</b>	<b>librarian</b>
<b>100</b>	<b>A</b>	<b>General Processing Data</b>
<b>101</b>	<b>A</b>	<b>Language of the Text, Soundtrack, etc.</b>
<b>105</b>	<b>A</b>	<b>Monograph Coded Data</b>

**Table 1: UNIMARC Default Subfields**



playing an actual role in live linking in Koha. Below is an example taken from our UNIMARC bibliographic records.

```
<datafield tag="481" ind1=" " ind2="1">  
  <subfield code="1">001295852</subfield>  
  <subfield code="a">Loura, Lu?s Armando de</subfield>  
  <subfield code="c">Lisboa. -[s.n.,</subfield>  
  <subfield code="d">D.L. 1956]</subfield>  
  <subfield code="t">Regula??o de avaria grossa pela arribada a Durban  
  
    do N. M. "Benguela", em 12 de Agosto de 1952</subfield>  
  
  <subfield code="5">PTBN: S.A. 5516//3 A.</subfield>  
</datafield>
```

## **Evaluation**

As we discussed in Section 3, UNIMARC has a more logical tag hierarchy ordered by function than MARC21, yet from a library systems point of view, MARC21 stands in a better position to be applied in the Koha system than UNIMARC and CMARC. This is because in MARC21, the most important coded data are designed in 007 and 008, and subfields in the two fields are all defined, while in UNIMARC and CMARC, the coded data spread into more than ten fields, and more than 90% of their subfields are not defined. Perhaps this explains the reason why the Koha system grows steadily in the US market more than in any other area in the world. From institutions' point of view, institutions which have sufficient technical support might not opt to apply a free integrated library system, but for institutions that do not have sufficient technical staff, and wish to apply free integrated library system, MARC21 could be the first option when considering Koha system. We mentioned in the Comparison, although all linking fields are not found in Koha default subfield, institutions can still redefine these accordingly depends on institution's practices.

## **4. Conclusion**

Koha is a mature integrated library system with good merits. Also, it is true that Koha provides default MARC21 and UNIMARC templates. This implies that Koha is designed rather to be used for MARC21 or UNIMARC but not for multi-scripts like CMARC, Japanese MARC or Korean MARC which need special programming. For countries with lower information technology development, enormous library system technical work is quite complex and requires in institutions wishing to do this a certain level of computer expertise which is not found in many developing countries.

Koha already has a substantial growing market in the US where MARC21 is the main format in use. In UK, although some libraries now use MARC21, some libraries still follow UKMARC which is more closely related to UNIMARC in the granularity of its fields, which probably is the main reason that Koha is only just starting there. In Taiwan, CMARC is closely intertwined to UNIMARC, and although Koha-Taiwan is volunteering putting ongoing effort to maintain and provide helpful assistance to those interested institutions, the Koha user community in Taiwan is small and not organized.

We conclude that implementing UNIMARC or MARC21 or even any other types of MARC formats in Koha along with a commercial-level-support like the business model in the US market should be our recommended solution.

### **References:**

- Breeding, Marshall, An Update on Open Source ILS, *Computers in Libraries*, 27(3): 27-29, 2007.
- Breeding, Marshall, The Viability of Open Source ILS, *ASIS&T Bulletin*, 35(2): 20-25, 2008.
- Mao, Ching-chen, Koha Taiwan, Google Groups, 2006,  
<http://groups.google.com/group/kohataiwan>, accessed at 20 March 2009.
- Mao, Ching-chen and Hsu, Ching-fen, Chinese MARC (Taiwan) and Its Bibliographic Database, *World Library and Information Congress: 72nd IFLA General Conference and Council 20-24 August 2006*, Seoul, Korea.
- Tsai, Yu-Chin, 2007, /trunk/cmarc/marc\_to\_utf8.pl - Koha Taiwan - Trac.  
[http://trac.koha-tw.org/koha/browser/trunk/cmarc/marc\\_to\\_utf8.pl](http://trac.koha-tw.org/koha/browser/trunk/cmarc/marc_to_utf8.pl).  
/trunk/cmarc/new\_split.pl - Koha Taiwan - Trac.  
[http://trac.koha-tw.org/koha/browser/trunk/cmarc/new\\_split.pl](http://trac.koha-tw.org/koha/browser/trunk/cmarc/new_split.pl), accessed at 20 March 2009.