Date : 21/07/2006

Japanese scripts and UNIMARC

Naoko HARAI
Director
Domestic Monographs Cataloging Division
Bibliography Department
National Diet Library, Japan

Meeting:	**77 UNIMARC**
Simultaneous Interpretation:	Yes

WORLD LIBRARY AND INFORMATION CONGRESS: 72ND IFLA GENERAL CONFERENCE AND COUNCIL
20-24 August 2006, Seoul, Korea
http://www.ifla.org/IV/ifla72/index.htm

Handling Japanese scripts in Japanese bibliographic data is quite important. I will introduce the characteristics of Japanese scripts and how we make bibliographic data in the Japanese language in UNIMARC format in consideration of these characteristics.

1. Kinds of Japanese script

Kinds of scripts used for writing Japanese

Kanji (Chinese characters and Chinese numerals) (ideograms): used for nouns, verb stems, etc.
Hira-kana (phonograms): particles, verb endings, etc.
Kata-kana (phonograms): foreign words, onomatopoeia, imitative words, etc.
Roman alphabet (Roman numeral): proper name, measure, abbreviation, etc.
Arabic numerals: numbers in general

In the Japanese writing system, the text is usually not written in a single script, as happens with the Roman alphabet but by a mixture of scripts such as kanji and hira-kana. In general, kanji (ideograms) are used to write nouns and verb stems, while hira-kana (phonogram) are used for particles and verb endings, that is, to make grammatical distinctions. Kata-kana (phonograms) are used for foreign words, onomatopoeia and imitative words. These scripts are used in combination. For numbers, we use kanji numerals, Roman numerals and Arabic numerals all together. Even the Roman alphabet is often used for proper names, measures and abbreviations in Japanese text.

Coded character set for Japanese language

	Single byte	Double-byte
Kanji	×	○
Hira-kana	×	○
Kata-kana	△	○

With these characteristics, what kinds of coded character sets are used to encode Japanese language as digital data? There are two kinds of scripts in the Japanese language; one can be rendered using single-byte code and the other requires double-byte code. For example, kanji cannot be encoded by single-byte code. Hira-kana and kata-kana are usually carried in double-byte code, while there are some single-byte character encoding schemes that support kata-kana. The Roman alphabet is commonly expressed by both single-byte and double-byte code.

Character encoding schemes for Japanese language

Single-byte	ASCII, EBCDIC
Double-byte	JIS code, Shift-JIS, EUC-JP
Other	UNICODE

JIS code is the most standard coded character set used in Japan. It is a double-byte set that supports several scripts such as kanji, hira-kana, kata-kana, Roman alphabet, Cyrillic alphabet and symbols. Shift-JIS and Japanese EUC (EUC-JP), which cover almost the same scripts as JIS code, are also widely used. For a single-byte encoding scheme, ASCII and EBCDIC are used. To express Japanese language as digital data, it is common for one double-byte encoding scheme to be used in combination with one

single-byte scheme.

However, such standard double-byte encoding schemes in Japan cover only Japanese kanji, Chinese characters used for Chinese language and hangul characters used for Korean are not included. In this regard, UNICODE that covers Japanese scripts, Chinese characters and hangul characters can express wider range of languages. And of course, UNICODE users have been increasing recently. For example, NACSIS-CAT, an online bibliographic database of the National Institute of Informatics (NII), uses UNICODE.

2. Reading of Japanese language

Reading of kanji		
Example 1		
原	gen	hara
原色	gensyoku	
原井		harai
原野	gennya	harano
Example 2		
花	ka	hana
華	ka	hana
華燭	kasyoku	

Next, I will explain the “reading” of the Japanese language. Kanji are ideograms and the reading of each character varies in Japanese. In general, one kanji has a number of readings, while most Chinese characters are read in a single way in Chinese and Korean. On the example 1, the kanji has two basic readings; “gen” and “hara”. The reading is decided in one way according to the word before or after it. In some contexts, it is not wrong to read one kanji in several ways.

There are some reverse cases, where several kanji can be used to express one reading. On the example 2, both kanji characters mean “flower” and their readings are also the same, “ka” and “hana”. However, according to context, there are some cases where both characters can be used and other where only one character is used.

Reading of personal names		
Example	東，幸雄	
Reading of family name		
	Higasi	Azuma
Reading of given name		
	Satio	Yukio

It is a common experience for us that when a person's name may have several readings we have to ask him/her how to pronounce it. As an example, I show the case that both family name and given name have several readings. This case shows that four kinds of combinations are possible. As an access point such as personal name, we have to specify not only its description such as kanji but also its reading. When the reading is different, even though they are written with the same kanji, the names are recognized to be different person's.

Reading in Japanese bibliographic data	
Example 1	
script unspecified	名称典拠のコントロール
kata-kana	メイショウ　テンキョ　ノ　コントロール
Roman alphabet	Meisyou tenkyo no kontororu
(description on material is only in unspecified script)	
Example 2	
script unspecified	東，幸雄
kata-kana	アズマ，ユキオ
Roman alphabet	Azuma，Yukio

In general, to make Japanese bibliographic data, first we describe as it appears on the material, that is, in unspecified mixture of scripts such as kanji and hira-kana. Then phoneticize the readings by kata-kana to make a heading. Data of readings are necessary as access points, whether they appear on the material or not, and they are entered as parallel data of those of unspecified scripts. Because the readings do not always appear on the material and we sometimes have to confirm them by research, it is time-consuming to enter reading data.

For headings, three kinds of data, unspecified scripts, kata-kana and Roman alphabet, are usually treated as one unit.

It is certain that these characteristics of the Japanese language mentioned above determine the form of Japanese bibliographic data. Next, I would like to explain how these characteristics are treated in bibliographic data.

3. JAPAN/MARC format

Characteristics of JAPAN/MARC format

A MARC format for Japanese established in 1980 by the NDL

- Non main entry system
  - Separation of description and access points
  - Title headings: 500-599
  - Author headings are alternative
- Treatment of Japanese scripts
  - Three scripts (scripts unspecified, kata-kana and Roman alphabet) treated as one unit in headings

When it started to input bibliographic data, the National Diet Library developed a MARC format that can treat the Japanese language appropriately. Because USMARC, the predecessor of MARC21, and UNIMARC did not have functions to treat non-Roman scripts at this time, we had to develop an original format. Reforming the UNIMARC format, the JAPAN/MARC format was developed to enable the Japanese language to be processed.

The main characteristics of JAPAN/MARC format are as follows:

First, the JAPAN/MARC format is not based on the main entry system. So the description fields (200-399) and the access point fields (500-799) are totally separated. All the title headings are recorded in the fields 500-599, without adopting the method using the indicator designating the title proper in the description fields as an access point. Author headings are alternative, priority level indicated by the order of entries, without distinction of main entry and added entry.

Secondly, the JAPAN/MARC format matches the characteristics of Japanese scripts. In principle, all the access point fields, that is, title headings, author headings and subject headings, treat data in three script types, script unspecified, kata-kana and Roman alphabet, as one unit. To create authority files, they are identified en bloc as a controlled name.

In addition, in the description and access point fields all the data are recorded in double-byte code, using JIS code. For the fields with single-byte code such as coded information, EBCDIC is used.

4. UNIMARC format

Although at the beginning the UNIMARC format was not suitable for languages not using the Roman alphabet, its later versions have been designed to treat bibliographic data in every language and every script in the world. I will explain how we deal with this format to record bibliographic data in the Japanese language in the NDL.

When we use Japanese scripts in UNIMARC format
- -To specify script of title proper
  Input “da” in 100 $a 34-35
- -For other scripts
  Data in kata-kana and Roman alphabet in fields such as headings
  Input “$7dc” or “$7ba” at the beginning of the field
- -To indicate the relationship between data of different scripts
  Set the same number in $6

First, we specify the script of title proper in the character position 34-35 of the field General Processing Data (100$a). For bibliographic data in Japanese, “da” which means “Japanese - script unspecified” should be input.

Then, if we have to record data in a different script from what has been coded there, we specify it again at the beginning of each field, in $7. This subfield is used when we input kata-kana and Roman alphabet to indicate “reading” for bibliographic data in Japanese. We input “$7dc” for kata-kana and “$7ba” for Roman alphabet at the beginning of the field.

When “da” or “dc” is input, JIS code is used as a character encoding scheme, and when “ba”, the scheme is ISO646 (IRV).

Also, we use $6 to indicate that the data in three types of scripts should be considered as one unit, as in JAPAN/MARC format. The same number set in $6 for the same tag which appears repeatedly indicates that the data with this number make one unit.

Example in title
200 1 $6a01$a 精神科救急
200 1 $6a01$7dc$a セイシンカ　キュウキュウ
200 1 $6a01$7ba$aSeisinka kyuukyuu
Example in Title and Statement of Responsibility
200 1 $6a01$a シェイクスピアの比喩研究$e 周辺劇作品を資料として$f 平岩紀夫　著
200 1 $6a01$7dc$a シェイクスピア　ノ　ヒユ　ケンキュウ
200 1 $6a01$7ba$aSyeikusupia no hiyu kenkyuu

Let me give you an example in Title. The title proper of this material is “Seisinka kyuukyuu”. Script is not specified in the first field because it has already been specified in 100 $a34-35 with “da.” The field 200 is repeatable only for inputting different scripts.

A major difference from JAPAN/MARC format is that the title headings are not totally separated from the description fields. For bibliographic data in European languages, it is sufficient to use the indicator for title proper to be designated as an access point. But for data in the Japanese language, we need to repeat the field of the same tag in order to include data on “reading” in addition to the use of the indicator. As a result, the number of data items which exist inside fields with the same tag differs, because we do not record “reading” for all data items other than title proper, for example, not for statement of responsibility.

Example of author headings
701 1$300085036$6a01$a 富岡$b 行昌$f１９２３－
701 1$6a01$7dc$a トミオカ，$b ユキマサ
701 1$6a01$7ba$aTomioka,$bYukimasa
701 1$300075508$6a02$a 鈴木$b 健二$f１９２９－$c 美学
701 1$6a02$7dc$a スズキ，$b ケンジ
701 1$6a02$7ba$aSuzuki,$bKenzi

The next example is of author headings. Fields with the same number in $6 make one unit, so when there are two authors, the record is like this. Additions are included only in the first field.

As the JAPAN/MARC uses the non main entry system, tag 701 is used for all the author headings of personal names. For author headings of corporate body names, 711 is used.

Example of a bibliographic record

```
(snip)
100   $a19890120d1986    m  y0jpnc0112    da
(snip)
200  1 $6a01$a 人間国宝中里無庵$e 炎の生涯$f 富岡行昌，鈴木健二　著
200  1 $6a01$7dc$a ニンゲン　コクホウ　ナカザト　ムアン
200  1 $6a01$7ba$aNingen kokuhou nakazato muan
210    $a 佐賀$c 佐賀新聞社$d１９８６．７
215    $7ba$a318p$d22cm
300    $a 中里無庵の肖像あり
320    $a 中里家年譜：ｐ２７９～２９９
600   1$300053200$6a01$a 中里$b 太郎右衛門$d１２世$f１８９５-$2ＮＤＬＳＨ
600   1$6a01$7dc$a ナカザト，$b タロウエモン
600   1$6a01$7ba$aNakazato,$bTarouemon
686    $7ba$aKB372$2NDLC
686    $7ba$a751.1$2NDC(8)
701   1$300085036$6a01$a 富岡$b 行昌$f１９２３-
701   1$6a01$7dc$a トミオカ，$b ユキマサ
701   1$6a01$7ba$aTomioka,$bYukimasa
701   1$300075508$6a02$a 鈴木$b 健二$f１９２９-$c 美学
701   1$6a02$7dc$a スズキ，$b ケンジ
701   1$6a02$7ba$aSuzuki,$bKenzi
(snip)
```

As above, the UNIMARC format is able to treat Japanese scripts although the method is different from that of the JAPAN/MARC format. Let us look at a bibliographic record as an example. All parts other than the general processing data, the description and the headings are omitted here.

Bibliographic data in Japanese mainly use script-unspecified fields, so we should be aware that data in kata-kana and Roman alphabet are given the status as “reading” and that not all elements are accompanied with reading. At present, items with reading in the description fields are title proper, parallel title, other title information, title proper of series, and publisher. Of titles of this example, only title proper has reading data.

5. Bibliographic Data in Japan

Situations of various MARCs in Japan	
Materials in Japanese language	JAPAN/MARC
	Other kinds of MARC
Materials in foreign languages	MARC21

JAPAN/MARC is merely one of the bibliographic data formats distributed in Japan. After the NDL launched JAPAN/MARC in 1981, several companies started to distribute bibliographic data in each distinctive format while they referred to JAPAN/MARC. Moreover, MARC 21 is the most popular format for bibliographic data of materials in foreign languages.

At large-scale libraries, it is common to create or provide bibliographic data in Japanese unique formats, including JAPAN/MARC, for Japanese books and in MARC21 format for foreign books.

The network of the NII, the representative bibliographic utility equipped with the function to create collaborative bibliography, maintains databases in its original format and also other libraries create and maintain bibliographic data in their own format.

Moreover, there are many of libraries using to the international bibliographic utilities such as the Online Computer Library Center (OCLC) and the Research Libraries Group (RLG).

It is true that the UNIMARC format existed as a basis for bibliographic data at the starting point in Japan, although it is unfortunately less common in Japan.

6. JAPAN/MARC in UNIMARC format

JAPAN/MARC in UNIMARC format	
1997	Launched JAPAN/MARC (A)
1998	Decided to develop bibliographic data in UNIMARC format
2000	Published “NDL CD-ROM Line: National Diet Library Catalog of Serials”
2003	Operated a new bibliographic data system
	Launched UNESCO’s Index Translationum in UNIMARC format

It is rational that the Japanese original formats including JAPAN/MARC are used within Japan. Needless to say, serious problems are caused by using a number of formats in parallel; for example, problems of compatibility among them and compatibility when bibliographic data created in Japan are distributed internationally.

The NDL recognized these problems and worked out countermeasures. UNIMARC format has been revised since JAPAN/MARC was developed, and consequently, it has become able to treat non-Roman alphabet data and is designed to deal with other kinds of resources. Accordingly, as a result of discussion, the NDL decided to develop

JAPAN/MARC in UNIMARC format in 1998. During the initial phases of the development, distribution in the forms of magnetic tape and CD-ROM was assumed; as for CD-ROM, the NDL published "NDL CD-ROM Line: National Diet Library Catalog of Serials" as of the end of 1999 and of 2000 in March 2000 and in March 2001 respectively.

The NDL thereafter reconstructed the master file of bibliographic data including review of the output design, and in 2003, the output program was completed. Since then the NDL has been providing UNESCO's Index Translationum in UNIMARC format. In addition, among the NDL bibliographic data, provision in UNIMARC format is possible for books, serials and non-book materials if requested.

The authority data have been developed using the UNIMARC format since the beginning of its distribution in 1997. The "JAPAN/MARC manual of authority 1st ed." was also published in 2003.