East Asian Language Support in OLIF2

Thomas Emerson
Basis Technology Corp.
One Kendall Square
Cambridge, MA 02139 USA
tree@basistech.com

December 10, 2001

1 Introduction

This document contains some initial thoughts and observations for encoding East Asian terminology resources using the OLIF2 DTD. It is based on the 2001 July 3 revision of “The Structure and Content of the Body of an OLIF v. 2 File”. It is very much a work in progress, and comments are solicited.

2 Features of East Asian Languages

2.1 Chinese

Chinese is written almost exclusively with ideographic characters, *hanzi* (汉字/漢字). Latin/Arabic numerals are used, as are some latin letters, generally in English words.

The People’s Republic of China (PRC) defined a phonetic alphabet based on the Latin script, 拼音 *pinyin*, where tones are marked using diacritics or numbers.

Taiwan’s phonetic alphabet, 註音符號 *Zhùyīn fúhào*, is based on greatly simplified forms of the *hanzi*. Zhuyin is also known as ”bopomofo”, after the first four sounds in the alphabet.

Chinese is often used as the canonical example of an isolating language: it has no inflectional morphology, though it does have a very rich and varied syntax.

Terminology rarely consists of loans; native words are coined representing the new concepts.

2.2 Japanese

Japanese makes use of four scripts:

- 漢字 *kanji*. Chinese characters used for nouns, verbs, proper names, etc.
- 平仮名/ひらがな *hiragana*. Syllabary used for grammatical function words, *kanji* readings and okurigana (verb/adjective conjugation).
- 片仮名/カタカナ *katakana*. Syllabary used for foreign words (including technical terms), mimetics, and for emphasis.
- ローマ字 *roomaji*. Latin/roman letters used for foreign words, emphasis, and “pop” words.

There are a number of romanization systems used for Japanese.

Technical terminology is often created through borrowing from English, with the term written in *katakana*. It is also common to see the English terms used directly in some contexts.

2.3 Korean

Korean is an agglutinative language. As in Japanese, there are a number of romanization schemes in use. This paper makes use of the Yale romanization.

Korean is written with two scripts:

- 한글 *hankul*. The native Korean alphabet, developed the 14th Century by the Sage King Sejong. Han'kul is used almost exclusively in written Korean.
- 한자 *hanca*. Chinese characters. In modern Korean *hanca* are used primarily for proper names.

Terminology is often created through borrowing from English or Japanese with the words adapted to the Korean sound system written in *hankul*. Technical terminology will almost never be written in *hanca*.

3 DTD Comments

3.1 <language>

The July OLIF specification states that the <language> element contains the ISO 639-1 language code for the language of the entry string. Unfortunately neither ISO 639-1 nor ISO 639-2 can encode the various topolects used in China. The same is true for ANSI/NISO Z39.53-1994.

SIL International's *Ethnologue* (Grimes 2000) assigns a unique three-letter code for each language in the world, including the Chinese topolects. Constable and Simons (2001) have investigated the merger of Ethnologue with ISO 639, though this work is not immediately applicable to OLIF.

RFC 3066 (Halvestrad 2001) provides a mechanism to support the level of identification required for full Chinese support, allowing second subtags of between three and eight characters to be registered with IANA. The IANA Language Tags registry (http://www.iana.org/assignments/language-tags) contains a set of second subtags for a number of Chinese topolects, including `zh-guoyu` for Mandarin and `zh-yue` for Cantonese.

Our recommendation is that OLIF use RFC 3066 for language identification instead of ISO 639-1.

3.2 <geogUsage>

Can this element be combined with the <language> element through the addition of the ISO 3166-1:1997 country code to the RFC 3066 language identifier? For example:

```
<language>zh-guoyu</language>
<geogUsage>HK</geogUsage>
```

and

```
<language>zh-guoyu-HK</language>
```

should be identical.

3.3 <orthVariant>

Should this element have an attribute indicating the type of the variant, similar to the <orthVariantType> element used in cross-references?

3.4 <orthVariantType>

Japanese terms may have several orthographic variants: katakana variants (length mark vs. extra vowel), roomaji, etc.

ATTRIBUTE	ATTRIBUTE (var.)	DESCRIPTION
japanese-1	jp-length-mark	Length mark vs. long vowel
japanese-2	jp-roomaji	Roomaji variant

There are orthographic differences between North and South Korea, though these need to be investigated further before they are codified.

The orthographic variation between the countries of Greater China is much more complex. As described in Section 2.1 there are two script variants used in written Chinese. In addition to these script differences, there are differences in character selection within a single script (e.g., Taiwan may use a variant full-form character than Hong Kong or the PRC).

There is no standard method for identifying the type of script being used for an entry. While ISO 15924 strives to provide this, like ISO 639-*x* it does not account for the differences in script used in written Chinese: it only gives the code "Hani" for Han Ideographs. However, ISO 15924 allows for private-use codes. This mechanism allows for the codes "Qjia" and "Qfan" to be used for simplified form and full-form characters, respectively.[1] Other scripts used for orthographic variants (the kana, han'gul, bopomofo, etc.) are represented unambiguously with ISO 15924.

Each of the East Asian languages has multiple romanization systems: the OLIF DTD does not provide an element or attribute to identify the romanization system for a given variant.

3.5 <PtOfSpeech>

There are several word classes used with East Asian languages that are not found in OLIF 2:

VALUE	DESCRIPTION
class	Classifier/Measure Word
verbn	Verbal Noun
vo	Verb-Object Compound
post	Postposition

4 Random Thoughts

The current OLIF DTD assumes left-to-right text directionality. In order to support the Semitic languages (esp. Arabic, Farsi, and Hebrew) directionality must be taken into account.

Within Chinese terminology the cross-reference groups may need to identify the language? I need to think about this further: the orthographic variation element(s) may provide all the information that is necessary.

Investigate the possibility of there being Asian-specific subject fields.

Define the <auxType> categories for CJK: *desu* for Japanese, *shi4* for Chinese, etc.

4.1 Speech Levels

Honorification is very important in Korean and Japanese. Korean has at least six levels, and Japanese three. The voice used when localizing content is very important and must be considered by the translator. Hence the speech level of a particular term/phrase must be represented. How is this handled for the EU languages currently encoded by OLIF (e.g., du/Sie or tu/vous). Are such discourse concerns applicable to this project?

5 References

Alvestrad, Harald. 2001. *Tags for the Identification of Languages*, BCP 47, RFC 3066, January 2001.

[1] Is it necessary to define private-use codes for simplified and full forms of Chinese ideographs? We believe it is: while section 4.1 of ISO 15924 shows a usage where one can concatenate the language and country codes to the script code. Unfortunately, this triple is not rich enough to unambiguously define the script. For example, a classification of "Hani zho CHN" for a document originating in China, written in "Chinese" (presumably Mandarin) with Han ideographs does not mean that the document is written in simplified characters. It could instead be written for Hong Kong, which uses full-form characters.

Constable, Peter and Gary Simons. 2001. *Mapping Between ISO 639 and SIL* Ethnologue*: Principles and Lessons Learned.* Dallas: SIL International.

Grimes, Barbara F. 2000. *Ethnologue*. 14th edition. 2 volumes. Dallas: SIL International. Web edition available online at http://www.ethnologue.com.

International Standards Organization. 1997. *ISO 3166-1:1997. Codes for the representation of names of countries and their subdivisions - Part 1: Country codes*. Geneva: International Organization for Standardization.

International Standards Organization. 1999. *ISO 12620:1999. Computer Applications in Terminology: Data Categories.* Geneva: International Organization for Standardization.

International Standards Organization. 2000. *ISO 15924:2000. Information Technology – Code for the representation of names of scripts*. Geneva: International Organization for Standardization.