Doc Type: Working Group Document
Title: Proposal to Update Syntax for Unicode/UCS Sequence Identifiers (USI) in ISO/IEC 10646
Source: U.S. National Body
Author: Ken Whistler
Status: National Body Contribution
Action: For consideration by JTC1/SC/WG2
Date: May 11, 2011

Introduction

Clause 6.6 of ISO/IEC 10646 defines the UCS Sequence Identifier (USI). The text of the clause in the FCD for the 3rd Edition currently reads as follows:

<quote>

ISO/IEC 10646 defines an identifier for any sequence of code points taken from the standard. Such an identifier is known as a UCS Sequence Identifier (USI). For a sequence of n code points it has the following form:

<UID1, UID2, ..., UIDn>

where UID1, UID2, etc. represent the short identifiers of the corresponding code points, in the same order as those code points appear in the sequence. If each of the code points in such a sequence has a character allocated to it, the USI can be used to identify the sequence of characters allocated at those code points. The syntax for UID1, UID2, etc. is specified in 6.5. A COMMA character (optionally followed by a SPACE character) separates the UIDs. The UCS Sequence Identifier includes at least two UIDs; it begins with a LESS-THAN SIGN and is terminated by a GREATER-THAN SIGN.

The full syntax of the notation of a UCS Sequence Identifier, in Backus-Naur form, is

"<" (xxxx | xxxxx | xxxxxx) ((",", space?) (xxxx | xxxxx | xxxxxx))+ ">"

where "x" represents one hexadecimal digit (0 to 9, A to F, or a to f).

</quote>

This notation specified in that clause follows widespread practice for citation of UCS character sequences in descriptive text. In such contexts, the use of angle brackets is not problematical, and in fact helps in visual identification of the sequences. The mix of commas and spaces also helps visually.

However, in data files, this notation is unnecessarily complicated to parse, and in actual practice, different, simpler notations are widely used in data files for the representation of UCS Sequences.

We propose to modify the text of Clause 6.6 to accomplish the following goals:

1. Make the specification of the syntax for UCS Sequence Identifiers (USI) clearer.

2. While retaining the validity of the existing definition, extend the allowed representation of the USI, so that formats widely implemented in data files will be recognized as valid USIs.

3. Make it simpler to maintain associated data files for specifying normative data such as the list of Named UCS Sequence Identifiers, without having to construct duplicate, parallel data files containing the same substantive content, but using distinct formats.

The revision for Clause 6.6 should use a more extended Backus-Naur form for the specification of the UCS Sequence Identifier (USI), so that it will be clear what is intended. As is already the case for the existing Clause 6.6, this specification makes use of the definition of UCS Short Identifiers (UID) from Clause 6.5.

<suggested_text_for_Clause_6_6>

ISO/IEC 10646 defines an identifier for any sequence of code points taken from the standard. Such an identifier is known as a UCS Sequence Identifier (USI).

The format of a USI depends on the definition of a UCS Short Identifier (UID), specified in Clause 6.5. The full format for a USI is specified by the following, in Backus-Naur form:

```
UCS_Sequence_Identifier := Unbracketed_Sequence | Bracketed_Sequence

Bracketed_Sequence := LEFTBRACKET Unbracketed_Sequence RIGHTBRACKET
Unbracketed_Sequence := Space_Delimited_Sequence | Comma_Delimited_Sequence
Space_Delimited_Sequence := UID (SPACE+ UID)+
Comma_Delimited_Sequence := UID (COMMA SPACE? UID)+

SPACE := U+0020
COMMA := U+002C
LEFTBRACKET := U+003C
RIGHTBRACKET := U+003E
```

In a UCS Sequence Identifier, the UID values occur in the same order as those code points appear in the sequence to be represented. If each of the code points in such a sequence has a character allocated to it, the USI can be used to identify the sequence of characters allocated at those code points. A UCS Sequence Identifier includes at least two UIDs.

Example 1. For typical use in descriptive text, or in printed tables meant to be read, a USI may be represented using a format which is more difficult to parse, but which facilitates reading. For

example, using a Bracketed_Sequence which contains a
Comma_Delimited_Sequence,
and which contains UIDs using the "U+" prefix:

   <U+0069, U+0307, U+0301>

Example 2. For typical use in data files, a USI may be represented using
a format which is easier for automatic parsing. For example, using an
Unbracketed_Sequence which contains a Space_Delimited Sequence,
and which contains UIDs without the "U+" or other prefixes:

   0069 0307 0301

</suggested_text_for_Clause_6_6>

If this change is adopted for the specification of the USI, then the
text of Clause 25 pertaining to the data file which defines Named
UCS Sequence Identifiers (NUSI) can also be simplified and modified
so that there will be no need to maintain multiple versions of
such data file with radically different syntax conventions.

Currently, the relevant text reads:

<quote>

The content linked to is a plain text file, using ISO/IEC 646-IRV
characters with LINE FEED as end of line mark, that specifies
after a 5-lines header, Named UCS Sequence Identifiers; each
line containing the following information organized in fields
delimited by a TAB character:

* 1st field: UCS sequence, following syntax defined in 6.6
* 2nd U : Name of the NUSI (following rules given in 23.5)

</quote>

We suggest that this be modified to the following text:

<suggested_text_for_Clause_25>

The content linked to is a plain text file, using ISO/IEC 646-IRV
characters with LINE FEED as end of line mark, that specifies
Named UCS Sequence Identifiers.

Each line in the text file contains the following information
organized in two fields:

* 1st field: Name of the NUSI (following the rules given in Clause 23.5)
* 2nd field: The USI associated with that Name (following the syntax
defined in Clause 6.6)

The two fields are delimited by a SEMICOLON (';') followed optionally
by zero or more SPACE characters. Comment lines, starting with a
NUMBER SIGN ('#') are informational only. Comment lines and blank
lines in the text file should be ignored by any automatic process
which parses the data file to extract the normative list of NUSIs.

</suggested_text_for_Clause_25>

The data file, NUSI.txt, should then be updated to use this revised specification from Clause 25. In particular, it should use the field order specified and use a SEMICOLON as the field delimiter, instead of a TAB character. (Note that use of a SEMICOLON as the explicit field delimiter eliminates potential parsing problems which can result from mixing of TAB and SPACE characters for delimitation.)

The revised data file should also mark the header lines explicitly with the comment line introduction character, so as to simplify the data parsing, and to bring it into line with the parsing already in widespread use for similar data files related to ISO/IEC 10646 content.