Title: Unicode Liaison Report to WG2

Date: 2012-2-13
Source: Unicode Consortium
Status: Liaison contribution
Action: For review by WG2 experts
Distribution: WG2

The Unicode Consortium is pleased to report on-going progress in development of the Universal Character Set resulting from collaboration with SC2, as well as progress on the Unicode Standard and related standards and technologies.

Publication of Unicode 6.1

Version 6.1 of the Unicode Standard was released on January 31, 2012. The character repertoire of this version is synchronized with ISO/IEC 10646 3rd edition. In addition to the encoding of new characters and scripts, some updates were made to data files affecting properties of previously encoded characters. As always, these changes are subject to Unicode stability policies.

UTC would like to point out a change to one pair of Unicode data files in particular, StandardizedVariants.txt and StandardizedVariants.html. These data files are not distributed with ISO/IEC 10646, but the content of StandardizedVariants.txt mirrors information listed in clause 16.5 of ISO/IEC 10646. For Unicode 6.1, UTC has added variation sequences for certain Emoji characters. Details are provided in a separate section of this report, below. This is mentioned here because of the impact on synchronization between standards: this change for Unicode 6.1 results in a minor point of divergence between Unicode 6.1 and the 3rd edition of ISO/IEC 10646.

Unicode Technical Reports normatively referenced in ISO/IEC 10646

At the Helsinki meetings of SC2, a resolution was taken (M17.04) to request that the Unicode Consortium adopt certain procedures in relation to updates to Unicode Technical Reports that are normatively referenced by ISO/IEC 10646. The Unicode Consortium acknowledges this request. It is the desire of the Consortium to continue constructive collaboration with SC2. To that end, the Consortium is happy to follow the procedures outlined as best as possible. We request that WG2 will keep the Consortium advised of which Unicode Technical Reports are normatively referenced in ISO/IEC 10646. It is our assumption that currently this includes the following:

- UAX #9 Unicode Bidirectional Algorithm
- UAX #15 Unicode Normalization Forms
- UTS #37 Ideographic Variation Database

Note that new versions for all three Technical Reports have recently been released. The updates to UAX #9 and UAX #15 did not contain any substantive, technical changes. The revision to UTS #37 is discussed further below.

Update of UTS#37, Ideographic Variation Database

At the Helsinki meeting of WG2, WG2 was notified of work in progress to update UTS#37. Version 3.1 of UTS#37 has since been published. (It has been posted in the WG2 document register as N4169; the current version is always available online at http://www.unicode.org/reports/tr37/.) The main changes in this new version pertain to the relationship between different registrations and are designed to improve the process for registering variation sequences and to better enable different agencies to collaborate in creating mutually-compatible and complementary registrations.

Emoji Variation Sequences

In response to requests from vendors implementing support for Emoji symbols encoded in Unicode 6.0 and ISO/IEC 10646:2011, UTC added new variation sequences in Unicode 6.1. These variation sequences are for certain characters from Japanese Emoji sets that were unified with existing UCS characters. Briefly, the problem encountered by implementers in relation to these characters is that two distinct presentations may be required, depending on the usage context, and that the presentation cannot be reliably predicted from the context. The requirement for implementers was to have a way to distinguish the desired presentation in plain text. This would, for example, allow one application to ensure that presentation was appropriate for Emoji usage scenarios, while another application could also ensure that presentation was appropriate for (say) dingbat presentation. After considering options, use of variation sequences was identified by UTC as an appropriate solution.

Therefore, the Unicode Consortium requests that WG2 accept variation sequences as described in N4182 for addition to the list of variation sequences in clause 16.5 of ISO/IEC 10646.

Proposed change to definition of *Ideographic description sequence*

Ideographic description characters and their use in ideographic description sequences (IDSs) are described in Annex I of ISO/IEC 10646. Annex I includes constraints on characters that can be used in an IDS; it also imposes certain length constraints for IDSs. The Unicode Standard has comparable text in chapter 12, “East Asian Scripts”.

The constraints on IDSs were originally added anticipating certain potential kinds of usage for IDSs in text interchange. After many years since IDSs were first defined, it is found that actual usage scenarios are more limited than anticipated: they are only used in the development and maintenance of the ISO/IEC 10646 and Unicode standards—specifically, in IRG processes and in UTR #45—and in similar scenarios of analyzing ideographs as candidates for unification or for distinct encoding. Moreover, it has been found that, for these usage scenarios, the current constraints on IDSs are too restrictive. In particular, there are instances in UTR #45 of complex ideographs requiring an IDS that exceeds the current length limitation (16 characters). Also, in IRG analysis of ideographs, private-use characters are sometimes used to represent an ideograph component that is useful in describing an ideograph but is itself not a candidate for encoding as an ideograph or radical.

In practice, the Annex I constraints are not providing any particular benefits for users or implementers: there are no software implementations at risk of breaking if the length limits or the constraint on use of PUA characters is not observed. Users may not be able to interpret IDS data containing PUA characters, but that is expected in any usage of PUA characters: interchange using PUA characters always assumes a private agreement between parties as to their interpretation. If such a private agreement exists, then there is no risk in interpretation of IDSs using those PUA charactesr.

While the Annex I constraints are not providing practical benefits, they are creating real problems. The requirement for UTR #45 and IRG to use IDSs violate Annex I constraints leave us open to criticism that our own data and processes do not conform to our own standards. UTC has already received this critique in relation to UTR #45. In view of this, it would make sense to eliminate those particular constraints that lead to such problems.

To that end, the Unicode Consortium requests that WG2 consider proposed changes to Annex I, as given in N4234. Additional background information on IDSs used by IRG is provided in N4241.

Process for encoding "urgently-needed" CJK ideographs

In 2007, IRG identified a small collection of "urgently-needed" characters. (These later became identified as *Extension D*.) Noting the lengthy times involved in processing large repertoires such as Extension A, B and C, IRG adopted a different process for these urgently-needed characters. Because they had identified a very small set, IRG work on these characters progressed very quickly, and they were encoded much faster than would have otherwise happened.

The Unicode Consortium sees the work on UNC / Extension D as having been very successful: important characters were encoded relatively quickly without sacrificing quality. This is beneficial for both implementers and end users of the UCS. Moreover, we believe that it would be of significant benefit for implementers and end users if a faster "UNC" process were repeated on a regular basis as other small sets of "urgently-needed" characters are identified.

To that end, the Unicode Consortium recommends to WG2 that it have IRG adopt a regular "UNC" process, as discussed in greater details in N4230.

Lithuanian text processing issues

The UTC considered documents submitted by the Lithuanian NB to WG2 (N4191, etc.). UTC endorses the stability policies adopted jointly by the Unicode Consortium and by ISO/IEC JTC1/SC2 that prevent encoding of additional pre-composed characters for the Latin script. At the same time, we recognize the legitimate concerns of the Lithuanian NB and Lithuanian users in relation to issues in processing of accented Lithuanian text. With that in mind, the UTC adopted the following resolution:

> UTC encourages its member companies to review their implementations to ensure the correct input and display of all Lithuanian characters.

Common Locale Data Repository (CLDR)

Unicode CLDR, Version 21, was released on February 10, 2012. CLDR 21 contains data for 193 languages and 170 territories: 528 locales in all. The focus of this release was on improvements to data formats, tools and processes used by the CLDR project, and consistency improvements in the data.

The Unicode Consortium feels confident that National Bodies and experts represented in WG2 will find the CLDR offers useful benefits in enabling support in software products for languages and cultures from across the world. As always, experts in WG2 are invited to participate in the on-going development of CLDR. Current information on CLDR can be found on the Unicode Web site at http://cldr.unicode.org/.