L2/13-125

Title: Unicode Liaison Report to WG2

Date: 2013-6-10
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.

Preparation of Unicode 6.3 for publication

Work has been in progress on version 6.3 of the Unicode Standard, with a target date for publication in September 2013.

The main focus for version 6.3 is updates to the Unicode Bidirectional Algorithm. These are described further below. The character repertoire for Unicode 6.3 will be synchronized with ISO/IEC 10646:2012, plus the addition of 5 layout control characters for use in bidirectional text that are among the additions in ISO/IEC 10646:2012/DAM2.

Preliminary content for Unicode 6.3 has been made publicly available for review and feedback. For additional information, see http://www.unicode.org/review/pri249/.

Unicode Technical Reports normatively referenced in ISO/IEC 10646

It is understood that ISO/IEC 10646 makes normative reference to these three specifications maintained by the Unicode Consortium:

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

Current versions for each of these are as follows:

- UAX #9, UAX #15: versions published as part of Unicode 6.2, September 2012. The 6.2 versions did not include any significant changes from the prior version.
- UTS #37: version 3.7, published November 7, 2011.

Updates to UAX #9 and UAX #15 are being prepared for Unicode 6.3. The changes to UAX #15 are minor. Version 6.3 of UAX #9, however, will involve significant changes. Details regarding the update to UAX #9 are described below.

Update of UAX#9, Unicode Bidirectional Algorithm (UBA)

There are two significant enhancements being made to the UBA to address significant problems that have been encountered by implementers in content publishing and localization scenarios.

The first of these involves addition of new layout control characters and associated changes to the algorithm. The new layout characters are used to indicate runs of text with a directional level embedding, similar to the existing embedding control characters U+202A LEFT-TO-RIGHT EMBEDDING and U+202B RIGHT-TO-LEFT EMBEDDING. The difference from the existing embedding controls has to do with the effect of characters surrounding the embedded run on the direction of the embedded run: with the existing embedding controls, surrounding characters significantly affect the directionality of the embedded run, but the new embedding controls isolate the embedded run from being affected by the surrounding characters. The new embedding layout characters, which are included in DAM2, are:

- U+2066 LEFT-TO-RIGHT ISOLATE
- U+2067 RIGHT-TO-LEFT ISOLATTE
- U+2068 FIRST STRONG ISOLATE
- U+2069 POP DIRECTIONAL ISOLATE

The second major enhancement has to do with paired punctuation characters with neutral directional properties. A common problem encountered with these characters is that a matching pair surrounding a run of text is not presented as would be expected. The cause is that, while intuitively a matching pair should be assigned to a directional level in a matching way, under the existing algorithm each is subject to local directionality effects and can be assigned mismatching directional levels. The enhancement to UBA will add logic to recognize matching paired punctuation in a string and to ensure that they are assigned a directional level in a consistent manner.

A draft of the update to UAX #9 has been submitted to the WG2 document register (N4446). Note that this is not the final draft; for more recent drafts, see the following URL:

http://www.unicode.org/reports/tr9/proposed.html

Chart glyphs for U+0145/U+0146 and requirements for Latvian and for Marshallese

A contribute recently received by the Unicode Technical Committee pointed out that U+0145 and U+0146 have names that reference CEDILLA as the combining mark, yet the glyphs that appear in the code charts show the combining mark COMMA BELOW. The comma form reflects the requirements for the Latvia language. However, the cedilla form is used for the Marshallese language. Some implementers have observed that they have difficulty supporting both language communities and have questioned whether *n with comma* and *n with cedilla* should be disunified. The Unicode Technical Committee have requested input on this issue from experts in Latvia, and would also welcome input from experts in WG2.

Proposed script name change: Kikakui

Among the additions in DAM 2 are characters for the “Mende” script. After considering expert input regarding this script, the Unicode Technical Committee recommends changing the name for this script to “Kikakui”.

On the script name "Hungarian"

Among the additions in DAM 2 are characters for the runic script referred to as "Hungarian", "Old Hungarian", "Szekely-Hungarian Rovas" or other names. In DAM 2, the name "Hungarian" is used in most of the document, though "Old Hungarian" appears in the title. The name used in the title and elsewhere should be the same.

The Unicode Technical Committee acknowledges that the name of this script has been a highly controversial matter since proposals were first considered within WG2 at the Dublin meeting in 2009. It has become clear that there is no ideal name that would be a strong preference for all stakeholders and that, for the script to be encoded, it will be necessary to adopt a compromise name that may not be a strong preference for any stakeholder. Noting that the term "Hungarian" occurs in the various names that have been acceptable to respective stakeholders, the Unicode Technical Committee concludes that that term may be the best available compromise name on which to establish consensus. Thus, we support the name "Hungarian" as used (apart from the title, as noted above) in DAM 2.

Lithuanian text processing issues

As reported previously, 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.

In addition, UTC communicated the Lithuanian concerns to the CLDR Technical Committee and requested that that TC consider actions that they might take to support the Lithuanian concerns. The CLDR TC did takes steps to provide information in the CLDR locale data for Lithuanian to help implementers understand requirements for Lithuanian text processing. An independent report with details on these additions has been submitted to WG2—see N4452.

Common Locale Data Repository (CLDR)

Unicode CLDR, Version 23.1, was released on May 15, 2013. CLDR 23.1 contains data for 215 languages and 227 territories—654 locales in all.

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/.