# Title: Unicode Liaison Report to SC2

Date: 2014-9-29
Source: Unicode Consortium
Status: Liaison contribution
Action: For review by SC2 members and WG2 experts
Distribution: SC2, 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 7.0

Version 7.0 of the Unicode Standard was published June 16, 2014. This version of the Unicode Standard is synchronized with ISO/IEC 10646:2012, plus Amendments 1 and 2. Additionally, it includes the accelerated publication of U+20BD RUBLE SIGN.

## 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 revision 31, published as part of Unicode 7.0, June 2014. This version has no substantive changes from the previous version that was published as part of Unicode 6.3.0.
- UAX #15 revision 41, published as part of Unicode 7.0, June 2014. This version includes only minor changes.
- UTS #37: version 3.1, published November 7, 2011.

As mentioned in a previous liaison report, software vendors that have begun to implement the Unicode 6.3 version of the Unicode Bidirectional Algorithm have identified some details in the revised specification that require revision. These will be prepared for a future version of the Unicode Standard to appear in 2015. WG2 members can expect a draft update to UAX #9 to be submitted for WG2 review in Q1 of 2015.

## Publication Schedule for the Unicode Standard

As mentioned in a previous liaison report, the Unicode Consortium will be moving to a regular release schedule for the Unicode Standard, with new versions to appear mid-year each year. Unicode 7.0 represents the beginning of this regular release schedule.

This publication schedule will have certain implications for synchronization with ISO/IEC 10646. The publication process for Unicode will require that the character repertoire be finalized by early February for a version to be released the following summer. In terms of process for 10646, characters can be included in a given version of Unicode if, by late January of that year, the characters are in a published edition or amendment of 10646 or are at least ready for approval stage (i.e., a DAM or DIS ballot has been completed and ballot comments have been resolved). In exceptional cases, such as new currency symbols, characters may be added to a given Unicode version that have not yet reached the approval stage in ISO process.

For example, if a repertoire for Amendment 1 has passed a DAM ballot with comments resolved at WG2 #63 in Colombo, then it will be possible to include that repertoire in Unicode's 2015 release. But, Amendment 2 would not be considered for Unicode's 2015 release since it would not be stabilized by January 2015.

WG2 may want to take Unicode's schedule into consideration when planning timelines for new editions and amendments. In particular, it will aide in synchronization if ballot resolution meetings for enquiry drafts (DIS, DAM) can be timed for Q4 of any given calendar year, or mid-January at the latest. For example, it would be desirable if a DAM ballot for Amendment 2 were completed with comments resolved by January 2016.

## Emoji Additions for Unicode 8.0

The large emoji sets that were added in Unicode 6.0 (synchronized with ISO/IEC 10646:2011) have since been implemented by several platform vendors with products used worldwide, gaining widespread adoption. The success of these implementations have not been without concern, however. In particular, there has been significant criticism of Unicode and of vendors implementing emoji due to a lack of ethnic diversity. This is a significant problem in certain regions such as North America and Europe in which there are significant populations of different racial backgrounds.

Due to these significant criticisms, members of the Unicode Consortium have an urgent need to address diversity concerns related to emoji encoded as part of Unicode 8.0, to be published in the summer of 2015.

A proposal relating to emoji diversity has been submitted for consideration at WG2 #63: N4599 Emoji Skin Tone Characters. The Unicode Consortium strongly requests that WG2 experts and SC2 members support this proposal or, at least, *promptly* provide feedback to Unicode on how to amend the proposal with a view to consensus. Also, in the coming months, Unicode experts will be considering other limited, additional emoji characters that address user concerns in relation to diversity.

As there will not be another WG2 or SC2 meeting before Unicode 8.0 is published, Unicode will of necessity have to define characters without the benefit of the SC2 balloting procedures. We suggest that this situation is comparable to what we sometimes encounter in relation to currency symbols. As with urgently-needed currency symbols that need to be implemented by vendors ahead of the process for amendments to 10646, we trust that SC2 will be willing to process such additions with a view to stability of implementations.

## New Tai Lue Encoding Model

The New Tai Lue script was encoded in ISO/IEC 10646:2003 and Unicode 4.0. This script is historically derived from Brahmi, the most immediate predecessor being the Tai Tham script as used for the Xishuangbanna Dai language. The New Tai Lue script was devised to be a simpler script than Tai Tham.

Because of its historic derivation from Brahmi, the encoding model originally assumed was the standard Indic encoding model in which character sequences reflect logical, reading order rather than visual order. This is reflected in 10646 primarily by the representative glyph of certain characters using a dotted circle "◌" to indicate their status as combining marks.

However, it is found that the primary user community has implemented the script using a visual encoding model. While some software vendors have implemented the script using the logical encoding model assumed in the standard, evidence reflects that these implementations are not in widespread usage.

As a result, the Unicode Technical Committee is evaluating a proposal to change the encoding model for New Tai Lue script from the logical encoding model used for most Indic scripts to the visual encoding model used for Thai, Lao and Tai Viet scripts. A public call for review and input has been published: Pubic Review Issue 281, Proposed encoding model change for New Tai Lue. UTC intends to make a decision on this proposal in time for the Unicode 8.0 release. Accordingly, input from WG2 experts is requested.

The impact on ISO/IEC 10646 of the proposed encoding model change would be that the code chart for the New Tai Lue block would have to be revised to change the representative glyph for the characters at code positions U+19B0 – U+19C0 and U+19C8 – U+19C9.

## Cyrillic Contractions in the Common Template Table (CTT) of ISO/IEC 14651

The CTT contains a number of Cyrillic contractions. Most of these are used for archaic forms or for smaller language communities. These contractions result in a 20 – 30% performance cost in collation implementations based on the CTT with impact on all languages using Cyrillic script, such as Russian, Ukrainian, etc. A majority of Cyrillic users would benefit if these low-usage contractions were removed from the CTT and implemented instead in language-specific tailorings.

For this reason, the Unicode Technical Committee recommends to SC2 that all Cyrillic contractions be removed from the CTT except for Й and й (U+0419 CYRILLIC CAPITAL LETTER SHORT I, U+0439 CYRILLIC SMALL LETTER SHORT I).

For additional details, see Unicode document L2/14-140 "CTT: Remove Most Cyrillic Contractions".

## WG2 Document Register

Over the past year, experts in the Unicode Consortium have encountered significant difficulty accessing documents in the WG2 document register hosted by DKUUG. On more than one occasion, the document register has been offline for several days at a time.

Also, it should be noted that most documents in the WG2 register regularly get mirrored in the document register of the Unicode Technical Committee. The UTC document register is a public register that has been in operation for over seventeen years with a dedicated maintenance staff.

With these points in mind, the Unicode Consortium offers to take over hosting of WG2 documents, in order to ensure a highly-reliable resource for WG2.

## Common Locale Data Repository (CLDR)

Unicode CLDR, Version 26, was released on September 18, 2014. This version has a significant increase in locale data over previous versions. The next version is being prepared for release in March, 2015.

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/.