Title: Unicode Liaison Report to WG2

Date: 2011-6-6
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 6.1

The Unicode Consortium is working on preparation of the next version of the Unicode Standard, version 6.1. This will be synchronized with ISO/IEC 10646 3rd edition. The planned schedule for release is Q1 of 2012. This plan assumes that the 3rd edition of ISO/IEC 10646 will go to FDIS ballot following the Helsinki meeting of WG2.

Format of Named UCS Sequence Identifiers in NUSI.txt

Both ISO/IEC 10646 and The Unicode Standard include data files for named UCS sequence identifiers. Different formats are used for the two files, however. This duplication of the same data in different formats adds to the cost of maintaining the two standards. Thus, it would be advantageous to converge on a common format. The US national body has submitted a proposal (N4063) aimed at arriving at a common format for this data file. The Unicode Consortium strongly supports the goal of having a common file format, and to that end endorses the US NB proposal.

The data file in ISO/IEC 10646 for NUSIs is NUSI.txt. Here's a historical summary regarding the format of this file:

Name UCS sequence identifiers (NUSIs) were introduced in the 2nd edition, based on the notion of named sequences that had previously been introduced into The Unicode Standard. In the CD for the 2nd edition, the defined sequences were given by reference to information published in the Unicode Standard. In response to CD ballot comments from Japan, his changed in the FCD for the 2nd edition with the introduction of the NUSI.txt data file. That data file was based on a data file already published as part of the Unicode Standard, NamedSequences.txt. However, ISO/IEC 10646 already had the notion of a UCS sequence identifier (USI) with a specified syntax. The FCD version of NUSI.txt did not give sequences using the syntax specified for USIs. In response to FCD ballot comments from Japan, the final format for NUSI.txt in the 2nd edition was changed to list sequences using the syntax defined for USIs.

The result, then, is that NUSI.txt in ISO/IEC 10646:2011 and NamedSequences.txt in The Unicode Standard have exactly the same information but use two different formats. In order to unify the formats, the US had a ballot comment on the CD for the 3rd edition of ISO/IEC 10646 (see US comment T5 in N3921) requesting that the syntax in NamedSequences.txt be used. This comment was withdrawn for procedural reasons since the project editor pointed out the conflict with earlier ballot comments from Japan. The goal of converging on file formats remains, however.

The US contribution in N4063 points out a valid distinction in usage for sequence identifiers in descriptive text intended to be read by humans versus in data files intended to be read by machines. But usage sequences are valid, but have different requirements. The suggestion, then, is that two formats be defined for sequence identifiers corresponding to the two usage scenarios. The Unicode Consortium supports this proposal. It will allow the formats of NUSI.txt and NamedSequences.txt to converge, and also provides a syntax for describing sequences appropriate for use in machine-readable data files generally.

Glyph change: U+2D7F TIFINAGH CONSONANT JOINER

Tifinagh Consonant Joiner, U+2D7F, is a virama-like character in that it suppresses an inherent vowel and causes a consonant-ligating behaviour. In terms of comparison with other virama-like characters, though, it is most like the Khmer “virama”, U+17D2 KHMER SIGN COENG, in that it has no traditional graphic form.

As in the case of U+17D2, users and implementers need some glyph that can be used in fallback situations so that the presence of this character in a text sequence is visible. Document N4069 is a request from the original proposers of U+2D7F requesting a glyph change for that character to a glyph that reflects the fallback presentation they would like to use.

The Unicode Consortium supports this request and intends to use this glyph in charts for the next version of The Unicode Standard. The Consortium requests that WG2 likewise approve this glyph change for ISO/IEC 10646.

Proposed character / script additions

Proposals have been submitted to WG2 for encoding Duployan (N3985), Wingdings and Webdings (N4022), and Tirhuta (N4035). Experts in the Unicode Consortium have reviewed these proposals and support their immediate inclusion in an amendment to ISO/IEC 10646.

In the case of Duployan, there was discussion of the original proposal at the Pusan meeting, with some disagreement on the encoding order. The different parties have discussed this issue and have reached a consensus position in N4088. The Unicode Consortium has not yet had opportunity to review that document. Nevertheless, the consensus process is supported, and thus the Consortium would be receptive to WG2 action in Helsinki in relation to Duployan based on this consensus document.

Ideograph Variation Database (IVD)

UTC has in the past brought to the attention of WG2 the existence of the Ideographic Variation Database (IVD) and encouraged WG2 and member bodies to consider the use of that mechanism as an alternative to encoding of additional compatibility ideographs. We welcomed the decision of Japan to begin utilizing this mechanism in the form of the Hanyo-Denshi collection in the IVD.

Many ideographs are still being processed by IRG. Of those ideographs, many can be considered glyph variants of already-encoded characters. Thus, it appears that the best treatment for many ideographs being processed by IRG may be to define IVD sequences for them rather than to encode new characters. In IRG process, however, once an ideograph is recognized as unifiable, it falls out of scope for further processing by IRG. Yet for some users, a need may remain to have an encoded representation that

selects those particular variants. Having done the work to identify the unification status of an ideograph, IRG would be in an ideal position to prepare proposals for IVD registration.

To that end, UTC recommends that the mandate for IRG be extended to include preparation of IVD registrations.

See document N4084 for additional discussion.

Proposed update to Unicode Technical Standard (UTS) #37, Ideographic Variation Database

The Unicode Consortium has approved the initiation of work to revise UTS#37, the document that establishes and specifies the procedures of the Ideographic Variation Database. The proposed update includes various revisions of potential interest to SC2 and WG2. These include certain clarifications and added constraints, such as a requirement that registrants must supply representative glyphs for variation sequences in the registration proposal. It also sanctions some additional actions a registrants can take, including the ability to supply additional representative glyphs for previously-registered sequences.

WG2 and IRG experts, SC members, or WG2 or SC2 as a body, are invited to review and submit comments on the proposed update. The working draft and an overview of the proposed changes can be found at http://www.unicode.org/review/pri184/index.html.

For details on the how to submit comments to the Unicode Consortium, see http://www.unicode.org/review/. (If WG2 or SC2 as a body choose to submit comments the Liaison can facilitate submission of those comments to the Unicode Consortium.)

Common Locale Data Repository (CLDR)

Unicode CLDR, Version 2.0, was released on May 25, 2011. CLDR 2.0 contains data for 200 languages and 183 territories: 618 locales in all. In this release, there was a concerted effort to flesh out modern coverage-level data for the top 55 languages, for an increase of over 45% more data fields in those languages

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://unicode.org/cldr/.