L2/10-378

Title: Unicode Liaison Report

Date: 2010-9-30
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.

Indian Rupee Sign

Noting the decision of the Government of India in July 2010 to adopt a new currency symbol to represent the India rupee, the Unicode Consortium supports the proposal from the India national body (N3887) to encode INDIAN RUPEE SIGN in the UCS at code position U+20B9.

Also, noting the urgent need to support this currency symbol that may be faced by information systems, software products, content providers, etc. throughout the world, and possibly also by other international standards, it is the desire of the Consortium to expedite the encoding of this character so as to allow implementers and users to begin using this character in the near future. To that end, the Consortium plans to include this character in version 6.0 of the Unicode Standard, and encourages WG2 and SC2 to include this character in the earliest-possible amendment or revision to ISO/IEC 10646.

Release of Unicode 6.0

The Consortium is in final stages of preparation for the release of Unicode 6.0, which is scheduled for October 11, 2010.

The content of Unicode 6.0 will be essentially synchronized with the new edition of ISO/IEC 10646 that was recently submitted to ITTF for FDIS ballot. We anticipate one small deviation from full synchronization in the India rupee sign: we plan to include the rupee sign in Unicode 6.0, whereas given the FDIS stage of balloting for the new edition it appears to be too late for the rupee sign to be included in the initial release of the next edition of ISO/IEC 10646.

While Unicode 6.0 is a major version release, we will not be publishing this version in hard copy as has been done for major versions in the past. The full text of Unicode 6.0 and all associated annexes, code charts and data files will be published online.

There have been significant editorial revisions to two normative annexes: UAX #15, *Unicode Normalization Forms*, and UAX #44, *Unicode Character Database*. There have also been small, clarifying changes to conformance clauses of UAX #9, *Unicode Bidi Algorithm*. There are no significant changes to conformance requirements, however.

The Unicode character properties for 6.0 include two new provisional properties: The **IndicMatraCategory** and **IndicSyllabicCategory** properties are intended to support processing Indic scripts and provide information regarding how characters behave in relation to formation of Indic syllable clusters. A new provisional data file, **Script_Extensions.txt**, is also being added to provide

additional information about characters used in multiple scripts. This information is intended for use with various types of processing, including font binding, data segmentation, regular expressions and spoof detection.

Feedback from WG2 experts or SC2 member bodies on these provisional data properties is welcomed.

For certain previously-encoded characters, some character properties will change in this release. Details on those property changes are provided in an appendix at the end of this report.

New named sequences and provisional name sequences will also be added in Unicode 6.0. Details are provided in an appendix at the end of this report.

Release of Unicode Collation Algorithm 6.0

The Consortium is also in final stages of preparation for the release of Version 6.0 of the Unicode Collation Algorithm (UCA). This update will be synchronized with version 6.0 of the Unicode Standard. Correspondingly, it will synchronize with the next editions ISO/IEC 10646 and ISO/IEC 14651 with the slight exception of the additional inclusion of the Indian rupee sign.

CJK Charts

CJK code charts published with Unicode 6.0 will use multi-column format for all CJK ideograph blocks, including CJK Extensions B, C and D as well as the CJK Compatibility Ideographs and CJK Compatibility Ideographs Supplement blocks. The Consortium appreciates the on-going work being done within IRG to review CJK multicolumn charts, and we look forward to quality improvements in multi-column charts for CJK Extension B that will be of benefit to the Unicode Standard as well as to the next edition of ISO/IEC 10646.

Ideograph Variation Database (IVD)

The Unicode Consortium received a joint submission from the Information Technology Standards Commission of Japan and Information Processing Society of Japan to register a collection of ideograph variation sequences: the “Hanyo-Denshi” collection. Per IVD process, this submission was published for a 90-day public review, which ended June 25, 2010. This was published as Public Review Issue 167, which can be viewed at http://www.unicode.org/ivd/pri/pri167/index.html. Feedback received was relayed to the submitters. The Consortium is waiting for the submitters to complete their assessment of feedback and any revisions they wish to make before posting the Hanyo-Denshi collection as an IVD registration.

The feedback received on the Hanyo-Denshi collection included information from Adobe regarding overlaps between the Hanyo-Denshi collection and the Adobe-Japan1 collection, which was registered as an IVD collection in late 2007. The IVD process allows different registrants to register different sequences for particular ideograph variants, giving registrants freedom to make independent judgments and to meet their particular needs. At the same time, it is recognized that overlapping registrations that use different sequences for variations shared between collections can create challenges to vendors and users who need to operate between different systems. Accordingly, we invite all IVD registrants to consider collaboration with other registrants so that, wherever possible, registrations may be complementary rather than competing. Also, the Consortium welcomes any suggestions for

improvements to the IVD (see UTS#37, Unicode Ideographic Variation Database for details regarding IVD process) that can facilitate greater complementarity in registrations.

UTS #46: Unicode IDNA Compatibility Processing

SC2 members are likely aware that new IETF specifications for internationalized domain names (IDNA2008) were finalized and published earlier this year. (The IDNA2008 specification is defined by a set of IETF RFCs. An informative overview is provided in http://tools.ietf.org/html/rfc5894.)

This new specification differs from the previous IDNA specification (IDNA2003) in various ways. The differences are mostly positive; in particular, the new version extends the character repertoire (IDNA2003 was previously restricted to a subset of characters in ISO/IEC 10646:2003 plus Amendment 1), and provides an automatic process for incorporating new characters that get added to the UCS. However, IDNA2008 does not maintain 100% backwards compatibility with IDNA2003:

- Some IDNs are valid in IDNA2003 but invalid in IDNA2008
- Some IDNs are valid in both IDNA2003 and IDNA2008 but can resolve to different destinations

These differences present potential for interoperability and security problems. Thus, it is clear that there are transitional challenges that may be faced by registrars, implementers and users.

In view of this, the Unicode Consortium has published *Unicode Technical Standard #46, Unicode IDNA Compatibility Processing*, which includes processing that can mitigate risks during a transitional period. This specification is targeted primarily at applications doing lookup of IDNs, but it also includes a strong recommendation for registries. The current version of this specification assumes Unicode version 5.2; an update for Unicode 6.0 is in the final stages of preparation and will be release shortly. It is available at http://www.unicode.org/reports/tr46/.

Common Locale Data Repository (CLDR)

CLDR, Version 1.8.1, was released on April 29 of this year; work is in progress on version 1.9, which is schedule for release in November. Version 1.9 will not include significant additions or changes to data. Rather, the focus for this release is on tooling and structure changes, such as the addition of new categories of locale 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://unicode.org/cldr/.

Comments on N3819, “Preliminary Proposal for Encoding Special Scripts and Characters in UCS for Uighur, Kazakh and Kirgiz”

At WG2 meeting 56, China submitted document N3819, requesting feedback from other national bodies. This has been reviewed by experts within the Unicode Consortium, and feedback has been provided in N3889. In summary, the assessment of the Unicode Consortium is that all of the text elements discussed in N3819 can be represented using characters or sequences of characters already encoded in the UCS. Please refer to N3889 for details.

Sinhala Numerals (N3888)

The Sri Lanka Standards Institution has submitted N3888, "Proposal to include Sinhala Numerals to the BMP and SMP of the UCS". Experts in the Unicode Consortium have reviewed this proposal and support the encoding of these characters in the UCS with the following changes recommended:

- 0DE6..0DEF: change names to SINHALA ASTROLOGICAL DIGIT ZERO, SINHALA ASTROLOGICAL DIGIT ONE, etc.
- In the names list for 0DE6..0DEF, use a shorter heading than that given in N3888 (e.g., "Sinhala Lith Illikam Digits", and shorter descriptive text, or remove the descriptive text altogether.

Appendix: Character Property Changes in Unicode 6.0

The following are character property changes to existing characters that will be made in Unicode 6.0:

Property: General_Category

Character	Old property value	New property value
06DE ARABIC START OF RUB EL HIZB	Me (Enclosing Mark)	So (Symbol, other)
0CF1 KANNADA SIGN JIHVAMULIYA	So (Symbol, other)	Lo (Letter, other)
0CF2 KANNADA SIGN UPADHMANIYA	So (Symbol, other)	Lo (Letter, other)
19DA NEW TAI LUE THAM DIGIT ONE	Nd (Number, decimal)	No (Number, other)
19DE NEW TAI LUE SIGN LAE	Po (Punctuation, other)	So (Symbol, other)
19DF NEW TAI LUE SIGN LAEV	Po (Punctuation, other)	So (Symbol, other)
2118 SCRIPT CAPITAL P	So (Symbol, other)	Sm (Symbol, math)

Property: Bidi_Class

Character	Old property value	New property value
06DE ARABIC START OF RUB EL HIZB	NSM (Nonspacing mark)	ON (Other neutral)
070F SYRIAC ABBREVIATION MARK	BN (Boundary neutral)	AN (Arabic number)
0CF1 KANNADA SIGN JIHVAMULIYA	ON (Other neutral)	L (Left to right)
0CF2 KANNADA SIGN UPADHMANIYA	ON (Other neutral)	L (Left to right)

Property: Joining_Type

Character	Old property value	New property value
06DE ARABIC START OF RUB EL HIZB	T (Transparent)	U (Non-joining)

Property: Joining_Group

Character	Old property value	New property value
06C3 ARABIC LETTER TEH MARBUTA GOAL	Hamza_On_Heh_Goal	Teh_Marbuta_Goal

Property: Line_Break

Character	Old property value	New property value
06DE ARABIC START OF RUB EL HIZB	CM (Combining mark)	AL (Alphabetic)
19DA NEW TAI LUE THAM DIGIT ONE	NU (Numeric)	SA (Complex context)

Property: Script

Character	Old property value	New property value
02EA MODIFIER LETTER YIN DEPARTING TONE MARK	Zyyy (Common)	Bopo (Bopomofo)
02EB MODIFIER LETTER YANG DEPARTING TONE MARK	Zyyy (Common)	Bopo (Bopomofo)
0600 ARABIC NUMBER SIGN	Zyyy (Common)	Arab (Arabic)
0601 ARABIC SIGN SANAH	Zyyy (Common)	Arab (Arabic)
0602 ARABIC FOOTNOTE MARKER	Zyyy (Common)	Arab (Arabic)
0603 ARABIC SIGN SAFHA	Zyyy (Common)	Arab (Arabic)
0CF1 KANNADA SIGN JIHVAMULIYA	Zyyy (Common)	Knda (Kannada)
0CF2 KANNADA SIGN UPADHMANIYA	Zyyy (Common)	Knda (Kannada)
302E HANGUL SINGLE DOT TONE MARK	Zinh (Inherited)	Hang (Hangul)
302F HANGUL DOUBLE DOT TONE MARK	Zinh (Inherited)	Hang (Hangul)

Property: Deprecated

Character	Old property value	New property value
0673 ARABIC LETTER ALEF WITH WAVY HAMZA BELOW	N (No)	Y (Yes)

Property: kRSUnicode

Character	Old property value	New property value
4E2C CJK UNIFIED IDEOGRAPH-4E2C	90.0	90.0 90'.0
8FB6 CJK UNIFIED IDEOGRAPH-8FB6	162.0	162.0 162'.0
9EC4 CJK UNIFIED IDEOGRAPH-9EC4	201.0	201.0 201'.0
22605 CJK UNIFIED IDEOGRAPH-22605	61.5 110.04	61.5 110.4
2305A CJK UNIFIED IDEOGRAPH-2305A	67.8	66.11

Property: kTotalStrokes

Character	Old property value	New property value
7D2B CJK UNIFIED IDEOGRAPH-7D2B	11	12

Property: Simple_Uppercase_Mapping

Character	Old property value	New property value
0265 LATIN SMALL LETTER TURNED H	(no mapping)	A78D LATIN CAPITAL LETTER TURNED H

Property: Simple_Titlecase_Mapping

Character	Old property value	New property value
0265 LATIN SMALL LETTER TURNED H	(no mapping)	A78D LATIN CAPITAL LETTER TURNED H

Property: Uppercase_Mapping

Character	Old property value	New property value
0265 LATIN SMALL LETTER TURNED H	(no mapping)	A78D LATIN CAPITAL LETTER TURNED H

Property: Titlecase_Mapping

Character	Old property value	New property value
0265 LATIN SMALL LETTER TURNED H	(no mapping)	A78D LATIN CAPITAL LETTER TURNED H

Changes were also made to some property values for the kIRG_*Source property to standardize syntax used in the property values.

Appendix: New Names Sequences and Provisional Named Sequences in Unicode 6.0

The following named sequences (named UCS sequence identifiers) will be added in Unicode 6.0:

Name	UCS Sequence
BENGALI LETTER KHINYA	0995 09CD 09B7
HIRAGANA LETTER BIDAKUON NGA	304B 309A
HIRAGANA LETTER BIDAKUON NGE	3051 309A
HIRAGANA LETTER BIDAKUON NGI	304D 309A
HIRAGANA LETTER BIDAKUON NGO	3053 309A
HIRAGANA LETTER BIDAKUON NGU	304F 309A
KATAKANA LETTER AINU CE	30BB 309A
KATAKANA LETTER AINU TO	30C8 309A
KATAKANA LETTER AINU TU	30C4 309A
KATAKANA LETTER BIDAKUON NGA	30AB 309A
KATAKANA LETTER BIDAKUON NGE	30B1 309A
KATAKANA LETTER BIDAKUON NGI	30AD 309A
KATAKANA LETTER BIDAKUON NGO	30B3 309A
KATAKANA LETTER BIDAKUON NGU	30AF 309A
LATIN SMALL LETTER AE WITH GRAVE	00E6 0300
LATIN SMALL LETTER HOOKED SCHWA WITH ACUTE	025A 0301
LATIN SMALL LETTER HOOKED SCHWA WITH GRAVE	025A 0300
LATIN SMALL LETTER OPEN O WITH ACUTE	0254 0301
LATIN SMALL LETTER OPEN O WITH GRAVE	0254 0300
LATIN SMALL LETTER SCHWA WITH ACUTE	0259 0301
LATIN SMALL LETTER SCHWA WITH GRAVE	0259 0300
LATIN SMALL LETTER TURNED V WITH ACUTE	028C 0301

Name	UCS Sequence
LATIN SMALL LETTER TURNED V WITH GRAVE	028C 0300
MODIFIER LETTER EXTRA-LOW EXTRA-HIGH CONTOUR TONE BAR	02E9 02E5

The following *provisional* named sequences (named UCS sequence identifiers) will be added in Unicode 6.0:

Name	UCS Sequence
SINHALA CONSONANT SIGN RAKAARAANSAYA	0DCA 200D 0DBB
SINHALA CONSONANT SIGN REPAYA	0DBB 0DCA 200D
SINHALA CONSONANT SIGN YANSAYA	0DCA 200D 0DBA