

Subject: UTC agenda item: Clarification of "decimal digits"

From: Asmus Freytag <asmusf@ix.netcom.com>

Date: Mon, 26 Jul 2010 14:06:38 -0700

To: Mistress of the Docs <l2doc@unicode.org>, Lisa Moore <lisam@us.ibm.com>

The specification of "decimal digit" in the standard is insufficiently precise.

Chapter 4.6 says:

 Decimal digits form a large subcategory of numbers consisting of those digits that can be used to form decimal-radix numbers. They include script-specific digits, but exclude characters such as Roman numerals and Greek acrophonic numerals. (Note that <1, 5> = 15 = fifteen, but <I, V> = IV = four.) Decimal digits also exclude the compatibility subscript or superscript digits to prevent simplistic parsers from misinterpreting their values in context.

Numbers other than decimal digits can be used in numerical expressions and may be interpreted by a numeric parser, but it is up to the implementation to determine such specialized uses.

What it needs to say is something like this:

 Decimal digits, as commonly understood, are digits that are used to form decimal-radix numbers. They include script-specific digits, but exclude characters such as Roman numerals and Greek acrophonic numerals. (Note that <1, 5> = 15 = fifteen, but <I, V> = IV = four.)

The general category=Nd and numeric_type=decimal are assigned to those characters that are both used as decimal digits and for which a full set of digits has been encoded, in ascending order of value, with digit 0 at the first code point.

Decimal digits, as defined by these property assignments, exclude some characters, such as the Han digits (see the first 10 entries in Table 4.xx "Primary Numeric Ideographs"), that serve both as letters and as digits and which are not encoded in sequence. They also exclude the compatibility subscript or superscript digits to prevent simplistic parsers from misinterpreting their values in context.

Numbers other than decimal digits can be used in numerical expressions and may be interpreted by a numeric parser, but it is up to the implementation to determine such specialized uses.

UAX#44 says:

 6) If the character has the property value Numeric_Type=Decimal, then the Numeric_Value of that digit is represented with an integer value in fields 6, 7, and 8. See the discussion of decimal digits in Chapter 4 in [Unicode].

Problem:

 There's currently no limit given to the numeric_value, it should be formally limited to 0..9

Also, UAX#44 should point out that nt=Decimal is only assigned to decimal digits that are encoded as a full sequence from 0->9.

WG2 Principles and Procedures:

 That document should be reviewed to ensure that it specifies that decimal digits are to be encoded contiguously and in order, with room left for missing digits so that, if digits are later used as part of a place-value notation (i.e. a decimal radix notation)

they can be used in that manner. Exceptions to be made only where (like numeric ideographs) the digits also serve as letters, or otherwise their use in decimal-radix notation can be safely excluded.

.