The Unicode Codepoints and IDNA

Patrik Fältström
paf@cisco.com

ICANN Meeting, Paris
2008-06-26
List of changes

• See http://stupid.domain.name/idnabis/
Abstract

• This document specifies rules for deciding whether a codepoint, considered in isolation, is a candidate for inclusion in an Internationalized Domain Name.
What is this

- This document reviews and classifies the collections of codepoints in the Unicode character set by examining various properties of the codepoints. It then defines an algorithm for determining a derived property value. It specifies a procedure and not a table of codepoints so that the algorithm can be used to determine code point sets independent of the version of Unicode that is in use.
• The list of codepoints that can be found in Appendix A is non-normative. Section 2 and Section 3 are normative.
Property values

- PROTOCOL VALID
- CONTEXTUAL RULE REQUIRED
- DISALLOWED
- UNASSIGNED

[Definition has changed between -00 and 01]
LetterDigits (A)

“Good codepoints”

- \texttt{generalCategory(cp)} is in \{Ll, Lu, Lo, Nd, Lm, Mn, Mc\}
LetterDigits (A)
“Good codepoints”

- Ll - Lowercase_Letter
- Lu - Uppercase_Letter
- Lo - Other_Letter
- Nd - Decimal_Number
- Lm - Modifier_Letter
- Mn - Nonspacing_Mark
- Mc - Spacing_Mark
Unstable (B)
Normalization and Casefolding

• $\text{toNFKC(toCaseFolded(toNFKC(cp)))} \neq \text{cp}$
IgnorableProperties (C)
Properties to ignore

- property(cp) is in {
  Default_Ignorable_Code_Point,
  White_Space,
  Noncharacter_Code_Point
}
IgnorableBlocks (D)
Blocks to ignore

- block(cp) in {
  Combining Diacritical Marks for Symbols,
  Musical Symbols,
  Ancient Greek Musical Notation,
  Private Use Area
}
LDH (E)
ASCII Letters, Digits, Hyphen

• cp is in {002D, 0030..0039, 0061..007A}
Exceptions (F)

Codepoints needing special treatment

• cp in {
    002D, 00B7, 02B9, 0375, 
    0483, 05F3, 05F4, 06FD, 
    06FE, 0F0B, 3005, 3007, 
    303B, 30FB
}
Category F
Exceptions
• 002D; CONTEXTO  # HYPHEN-MINUS
• 00B7; CONTEXTO  # MIDDLE DOT
• 02B9; CONTEXTO  # MODIFIER LETTER PRIME
• 0375; CONTEXTO  # GREEK LOWER NUMERAL SIGN (KERAIA)
• 0483; CONTEXTO  # COMBINING CYRILLIC TITLO
• 05F3; CONTEXTO  # HEBREW PUNCTUATION GERESH
• 05F4; CONTEXTO  # HEBREW PUNCTUATION GERSHAYIM
• 06FD; PVALID  # ARABIC SIGN SINDHI AMPERSAND
• 06FE; PVALID  # ARABIC SIGN SINDHI POSTPOSITION MEN
• 0F0B; PVALID  # TIBETAN MARK INTERSYLLABIC TSHEG
• 3005; CONTEXTO  # IDEOGRAPHIC ITERATION MARK
• 3007; PVALID  # IDEOGRAPHIC NUMBER ZERO
• 303B; CONTEXTO  # VERTICAL IDEOGRAPHIC ITERATION MARK
• 30FB; CONTEXTO  # KATAKANA MIDDLE DOT
BackwardCompatible (G)
Backward compatibility

• cp in {}
JoinControl (H)

• property(cp) is in { Join_Control
Unassigned (J)
Unassigned codepoints

This definition has changed from version -00!

• cp is in \{Cn\} and property(cp) is not in \{Noncharacter_Code_Point\}
Algorithm

This definition has changed from version -00!

- Exceptions, see table
- BackwardCompatible, see table
- Unassigned, UNASSIGNED
- LDH, PVALID
- JoinControl, CONTEXTJ
- Unstable, DISALLOWED
- IgnorableProperties, DISALLOWED
- IgnorableBlocks, DISALLOWED
- LetterDigits, PVALID
- not LetterDigits, DISALLOWED

Exceptions
Backward compatibility
Unassigned codepoints
ASCII LDH
Special for lookup, resolution
Normalization, Casefolding
Properties to ignore
Blocks to ignore
Good codepoints
“The rest”
Remember that table is non-normative!

paf@cisco.com