The Unicode Codepoints and IDNA

draft-ietf-idnabis-tables-01.txt

Patrik Fältström
paf@cisco.com

ICANN Meeting, Paris
2008-06-26

List of changes

- See http://stupid.domain.name/idnabis/

skipping to change at *page 4, line 13*

others, requires that it not be used in labels unless specific other characters or properties are present. The abbreviated term CONTEXT is used to refer to this value in the rest of this document. There are two subdivisions of CONTEXTUAL RULE REQUIRED, one for Join_controls (called CONTEXTJ) and and for other characters (called CONTEXTO). These are discussed in more detail below and in [IDNA2008-protocol].

o DISALLOWED: Those that should clearly not be included in IDNs. Codepoints with this property value will never be permitted in IDNs.

o UNASSIGNED: Those codepoints that are unassigned in the Unicode Standard.

The (non-normative) table in Appendix A is derived from data in Unicode 5.1, rather than the earlier Unicode 3.2; this in order to take advantage of the expanded character repertoire and better definitions in the newer version. The mechanisms described here allow determination of the value of the property for future versions of Unicode (including characters added after Unicode 5.1). It should be suitable for newer revisions of Unicode, as long as the Unicode properties on which it is based remain stable.

skipping to change at *page 7, line 47*

H: property(cp) is in {Join_Control}

This category consists of Join Control characters (i.e., they are not in LetterDigits (Section 2.1.1)) but are still required in IDN labels under some circumstances. They require extended special treatment in Lookup and Resolution.

2.2.5. Unassigned (J)

J: cp is unassigned

This category consists of codepoints in the Unicode character set that are not (yet) assigned.

skipping to change at *page 4, line 13*

others, requires that it not be used in labels unless specific other characters or properties are present. The abbreviated term CONTEXT is used to refer to this value in the rest of this document. There are two subdivisions of CONTEXTUAL RULE REQUIRED, one for Join_controls (called CONTEXTJ) and and for other characters (called CONTEXTO). These are discussed in more detail below and in [IDNA2008-protocol].

o DISALLOWED: Those that should clearly not be included in IDNs. Codepoints with this property value will never be permitted in IDNs.

o UNASSIGNED: Those codepoints that are not designated (unassigned) in the Unicode Standard.

The (non-normative) table in Appendix A is derived from data in Unicode 5.1, rather than the earlier Unicode 3.2; this in order to take advantage of the expanded character repertoire and better definitions in the newer version. The mechanisms described here allow determination of the value of the property for future versions of Unicode (including characters added after Unicode 5.1). It should be suitable for newer revisions of Unicode, as long as the Unicode properties on which it is based remain stable.

skipping to change at *page 7, line 47*

H: property(cp) is in {Join_Control}

This category consists of Join Control characters (i.e., they are not in LetterDigits (Section 2.1.1)) but are still required in IDN labels under some circumstances. They require extended special treatment in Lookup and Resolution.

2.2.5. Unassigned (J)

J: cp is in {Cn} and property(cp) is not in {Noncharacter_Code_Point}

This category consists of codepoints in the Unicode character set that are not (yet) assigned. It should be noted that the set of

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.

Algoritm / table

- 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”

- `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

- `toNFKC(toCaseFolded(toNFKC(cp))) != 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)

Require extended special treatment in Lookup and Resolution

- 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 — Exceptions
- BackwardCompatible, see table — Backward compatibility
- Unassigned, UNASSIGNED — Unassigned codepoints
- LDH, PVALID — ASCII LDH
- JoinControl, CONTEXTJ — Special for lookup, resolution
- Unstable, DISALLOWED — Normalization, Casefolding
- IgnorableProperties, DISALLOWED — Properties to ignore
- IgnorableBlocks, DISALLOWED — Blocks to ignore
- LetterDigits, PVALID — Good codepoints
- not LetterDigits, DISALLOWED — “The rest”

Remember that table is non-normative!

paf@cisco.com