Network Working Group
Internet-Draft
Obsoletes: 3490 (if approved)
Intended status: Standards Track
Expires: November 22, 2008

J. Klensin, Ed.

May 21, 2008

Internationalizing Domain Names in Applications (IDNA): Protocol draft-ietf-idnabis-protocol-00.txt

Status of this Memo

This Internet-Draft will expire on November 22, 2008.

Abstract

This document supplies the protocol definition for a revised and updated specification for internationalized domain names (IDNs). The rationale for these changes, the relationship to the older specification, and important terminology are provided in other documents. This document specifies the protocol mechanism, called Internationalizing Domain Names in Applications (IDNA), for registering and looking up IDNs in a way that does not require changes to the DNS itself. IDNA is only meant for processing domain names, not free text.

Table of Contents

1. Introduction

This document supplies the protocol definition for a revised and updated specification for internationalized domain names. The rationale for these changes and relationship to the older specification and some new terminology is provided in other documents, notably [IDNA2008-Rationale].

IDNA works by allowing applications to use certain ASCII string labels (beginning with a special prefix) to represent non-ASCII name labels. Lower-layer protocols need not be aware of this; therefore IDNA does not depend on changes to any infrastructure. In particular, IDNA does not depend on any changes to DNS servers, resolvers, or protocol elements, because the ASCII name service provided by the existing DNS is entirely sufficient for IDNA.

IDNA is applied only to DNS labels. Standards for combining labels into fully-qualified domain names and parsing labels out of those names are covered in the base DNS standards [RFC1035]. An application may, of course, apply locally-appropriate conventions to the presentation forms of domain names as discussed in [IDNA2008-Rationale].

While they share terminology, reference data, and some operations, this document describes two separate protocols, one for IDN registration (Section 4) and one for IDN lookup (Section 5).

A good deal of the background material that appeared in RFC 3490 has been removed from this update. That material is either of historical interest only or has been covered from a more recent perspective in RFC 4690 [RFC4690] and [IDNA2008-Rationale].

[[anchor2: Note in Draft: This document still needs more specifics about how to perform some of the tests in the Registration and Lookup protocols described below. Those details will be supplied in a later revision, but the intent should be clear from the existing text.]]

1.1. Discussion Forum

[[anchor4: RFC Editor: please remove this section.]]

This work is being discussed in the IETF IDNABIS WG and on the mailing list idna-update@alvestrand.no

2. Terminology

General terminology applicable to IDNA, but with meanings familiar to

those who have worked with Unicode or other character set standards and the DNS, appears in [IDNA2008-Rationale]. Terminology that is an integral, normative, part of the IDNA definition, including the definitions of "ACE", appears in that document as well. Familiarity with the terminology materials in that document is assumed for reading this one. The reader of this document is assumed to be familiar with DNS-specific terminology as defined in RFC 1034 [RFC1034].

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14, RFC 2119 [RFC2119].

3. Requirements and Applicability

3.1. Requirements

IDNA conformance means adherence to the following requirements:

1. Whenever a domain name is put into an IDN-unaware domain name slot (see Section 2 and [IDNA2008-Rationale]), it MUST contain only ASCII characters (i.e., must be either an A-label or an LDH-label), or must be a label associated with a DNS application that is not subject to either IDNA or the historical recommendations for "hostname"-style names [RFC1034].

2. Comparison of labels MUST be done on the A-label form, using an ASCII case-insensitive comparison as with all comparisons of DNS labels.

3. Labels being registered MUST conform to the requirements of Section 4. Labels being looked up and the lookup process MUST conform to the requirements of Section 5.

3.2. Applicability

IDNA is applicable to all domain names in all domain name slots except where it is explicitly excluded. It is not applicable to domain name slots which do not use the LDH syntax rules.

This implies that IDNA is applicable to many protocols that predate IDNA. Note that IDNs occupying domain name slots in those older protocols MUST be in A-label form until and unless those protocols and implementations of them are upgraded.

3.2.1. DNS Resource Records

IDNA applies only to domain names in the NAME and RDATA fields of DNS resource records whose CLASS is IN.

There are currently no other exclusions on the applicability of IDNA to DNS resource records. Applicability depends entirely on the CLASS, and not on the TYPE except as noted below. This will remain true, even as new types are defined, unless there is a compelling reason for a new type that requires type-specific rules. The special naming conventions applicable to SRV records are examples of type-specific rules that are incompatible with IDNA coding. Hence the label on a record with TYPE SRV MUST NOT be an A-label or U-label (while it would be possible to write a non-ASCII string with a leading underscore, conversion to an A-label would be impossible without loss of information and because the underscore is not a letter, digit, or hyphen.

3.2.2. Non-domain-name Data Types Stored in the DNS

Although IDNA enables the representation of non-ASCII characters in domain names, that does not imply that IDNA enables the representation of non-ASCII characters in other data types that are stored in domain names, specifically in the RDATA field for types that have structured RDATA format. For example, an email address local part is stored in a domain name in the RNAME field as part of the RDATA of an SOA record (hostmaster@example.com would be represented as hostmaster.example.com). IDNA specifically does not update the existing email standards, which allow only ASCII characters in local parts. Even though work is in progress to define internationalization for email addresses [RFC4952], changes to the email address part of the SOA RDATA would require action in other standards, specifically those that specify the format of the SOA RR.

4. Registration Protocol

This section defines the procedure for registering an IDN. The procedure is implementation independent; any sequence of steps that produces exactly the same result for all labels is considered a valid implementation.

4.1. Proposed label

The registrant submits a request for an IDN. The user typically produces the request string by the keyboard entry of a character sequence in the local native character set. The registry MAY permit submission of labels in A-label form. If it does so, it SHOULD

perform a conversion to a U-label, perform the steps and tests described below, and verify that the A-label produced by the step in Section 4.5 matches the one provided as input. If, for some reason, it does not, the registration MUST be rejected.

4.2. Conversion to Unicode and Normalization

Some system routine, or a localized front-end to the IDNA process, ensures that the proposed label is a Unicode string. That string MUST be in Unicode Normalization Form C (NFC [Unicode-UAX15]).

As a local implementation choice, the implementation MAY choose to map some forbidden characters to permitted characters (for instance mapping uppercase characters to lowercase ones), displaying the result to the user, and allowing processing to continue. However, it is strongly recommended that, to avoid any possible ambiguity, entities responsible for zone files ("registries") accept registrations only for A-labels (to be converted to U-labels by the registry) or U-labels actually produced from A-labels, not forms expected to be converted by some other process.

4.3. Permitted Character and Label Validation

4.3.1. Rejection of Characters that are not Permitted

The Unicode string is examined to prohibit characters that IDNA does not permit in input. Those characters are identified in the "DISALLOWED" and "UNASSIGNED" lists that are discussed in [IDNA2008-Rationale]. The normative rules for producing that list and the initial version of it are specified in [IDNA2008-Tables]. Characters that are either DISALLOWED or UNASSIGNED MUST NOT be part of labels being processed for registration in the DNS.

4.3.2. Label Validation

The proposed label (in the form of a Unicode string, i.e., a putative U-label) is then examined, performing tests that require examination of more than one character.

4.3.2.1. Rejection of Confusing or Hostile Sequences in U-labels

The Unicode string MUST NOT contain "--" (two consecutive hyphens) in the third and fourth character positions.

4.3.2.2. Leading Combining Marks

The first character of the string is examined to verify that it is not a combining mark. If it is a combining mark, the string MUST NOT

be registered.

4.3.2.3. Contextual Rules

Each code point is checked for its identification as characters requiring contextual processign for registration (the list of characters appears as the combination of CONTEXTJ and CONTEXTO in [IDNA2008-Tables]). If that indication appears, the table of contextual rules is checked for a rule for that character. If no rule is found, the proposed label is rejected and MUST NOT be installed in a zone file. If one is found, it is applied (typically as a test on the entire label or on adjacent characters). If the application of the rule does not conclude that the character is valid in context, the proposed label MUST BE rejected. (See the IANA Considerations: IDNA Context Registry section of [IDNA2008-Rationale] and Appendix A of this document.)

4.3.2.4. Labels Containing Characters Written Right to Left

Additional special tests for right-to-left strings are applied (See [IDNA2008-BIDI]. Strings that contain right to left characters that do not conform to the rule(s) identified there MUST NOT be inserted in zone files.

4.3.3. Registration Validation Summary

Strings that have been produced by the steps above, and whose contents pass the above tests, are U-labels.

To summarize, tests are made here for invalid characters, invalid combinations of characters, and for labels that are invalid even if the characters they contain are valid individually. For example, labels containing invisible ("zero-width") characters may be permitted in context with characters whose presentation forms are significantly changed by the presence or absence of the zero-width characters, while other labels in which zero-width characters appear may be rejected.

4.4. Registry Restrictions

Registries at all levels of the DNS, not just the top level, are expected to establish policies about the labels that may be registered, and for the processes associated with that action. While exact policies are not specified as part of IDNA2008 and it is expected that different registries may specify different policies, there SHOULD be policies. These per-registry policies and restrictions are an essential element of the IDNA registration protocol even for registries (and corresponding zone files) deep in

the DNS hierarchy. As discussed in [IDNA2008-Rationale], such restrictions have always existed in the DNS.

The string produced by the above steps is checked and processed as appropriate to local registry restrictions. Application of those registry restrictions may result in the rejection of some labels or the application of special restrictions to others.

4.5. Punycode Conversion

The resulting U-label is converted to an A-label (i.e., the encoding of that label according to the Punycode algorithm [RFC3492] with the prefix included, i.e., the "xn--..." form).

4.6. Insertion in the Zone

The A-label is registered in the DNS by insertion into a zone.

5. Domain Name Resolution (Lookup) Protocol

Resolution is conceptually different from registration and different tests are applied on the client. Although some validity checks are necessary to avoid serious problems with the protocol (see Section 5.5 ff.), the resolution-side tests are more permissive and rely heavily on the assumption that names that are present in the DNS are valid. Among other things, this distinction, applied carefully, facilitates expansion of the permitted character lists to include new scripts and accommodate new versions of Unicode without introducing ambiguity into domain name processing.

5.1. Label String Input

The user supplies a string in the local character set, typically by typing it or clicking on, or copying and pasting, a resource identifier, e.g., a URI [RFC3986] or IRI [RFC3987] from which the domain name is extracted. Or some process not directly involving the user may read the string from a file or obtain it in some other way. Processing in this step and the next two are local matters, to be accomplished prior to actual invocation of IDNA, but at least these two steps must be accomplished in some way.

5.2. Conversion to Unicode

The local character set, character coding conventions, and, as necessary, display and presentation conventions, are converted to Unicode (without surrogates), paralleling the process described above in Section 4.2.

5.3. Character Changes in Preprocessing or the User Interface

The Unicode string MAY then be processed, in a way specific to the local environment, to make the result of the IDNA processing match user expectations. For instance, it would be reasonable, at this step, to convert all upper case characters to lower case, if this makes sense in the user's environment.

Other examples of processing for localization might be applied, if appropriate, at this point. They include interpreting the KANA MIDDLE DOT as separating domain name components from each other, mapping different "width" forms of the same character into the one form permitted in labels, or giving special treatment to characters whose presentation forms are dependent only on placement in the label. Such localization changes are even further outside the scope of this specification than the ones mentioned above.

Recommendations for preprocessing for global contexts (i.e., when local considerations do not apply or cannot be used) and for maximum interoperability with labels that might have been specified under liberal readings of IDNA2003 are given in [IDNA2008-Rationale].

Because these transformations are local, it is important that domain names that might be passed between systems (e.g., in IRIs) be U-labels or A-labels and not forms that might be accepted locally as a consequence of this step. This step is not standardized as part of IDNA, and is not further specified here.

5.4. A-label Input

If the input to this procedure appears to be an A-label (i.e., it starts in "xn--"), the lookup application MAY attempt to convert it to a U-label and apply the tests of Section 5.5 and, of course, the conversion of Section 5.6 to that form. If the A-label is converted to a U-label then the processing specified in those two sections MUST yield an A-label identical to the original one. See also Section 6.1.

In general, that conversion and testing should be performed if the domain name will later be presented to the user in native character form (this requires that the lookup application be IDNA-aware). Applications that are not IDNA-aware will obviously omit that testing; others may treat the string as opaque to avoid the additional processing at the expense of providing less protection and information to users.

5.5. Validation and Character List Testing

In parallel with the registration procedure, the Unicode string is checked to verify that all characters that appear in it are valid for IDNA resolution input. As discussed above and in [IDNA2008-Rationale], the resolution check is more liberal than the registration one. Putative labels with any of the following characteristics MUST BE rejected prior to DNS lookup:

- Labels containing code points that are unassigned in the version of Unicode being used by the application, i.e., in the "Unassigned" Unicode category or the UNASSIGNED category of [IDNA2008-Tables].

- Labels that are not in NFC form.

- Labels containing prohibited code points, i.e., those that are assigned to the "DISALLOWED" category in the permitted character table [IDNA2008-Tables].

- Labels containing code points that are shown in the permitted character table as requiring a contextual rule and that are flagged as requiring exceptional special processing on lookup ("CONTEXTJ" in the Tables) MUST conform to the rule, which MUST be present.

- Labels containing other code points that are shown in the permitted character table as requiring a contextual rule ("CONTEXTO" in the tables), but for which no such rule appears in the table of rules. With the exception in the rule immediately above, applications resolving DNS names or carrying out equivalent operations are not required to test contextual rules, only to verify that a rule exists.

- Labels whose first character is a combining mark. [[anchor17: Note in Draft: this definition may need to be further tightened.]]

In addition, the application SHOULD apply the following test. The test may be omitted in special circumstances, such as when the resolver application knows that the conditions are enforced elsewhere, because an attempt to resolve such strings will almost certainly lead to a DNS lookup failure. However, applying the test is likely to give much better information about the reason for a lookup failure -- information that may be usefully passed to the user when that is feasible -- then DNS resolution failure alone.

- Verification that the string is compliant with the requirements for right to left characters, specified in [IDNA2008-BIDI].

For all other strings, the resolver MUST rely on the presence or absence of labels in the DNS to determine the validity of those labels and the validity of the characters they contain. If they are registered, they are presumed to be valid; if they are not, their possible validity is not relevant. A resolver that declines to look up a string that conforms to the above rules is not in conformance with this protocol.

5.6. Punycode Conversion

The validated string, a U-label, is converted to an A-label using the punycode algorithm.

5.7. DNS Name Resolution

The A-label is looked up in the DNS, using normal DNS procedures.

6. Name server Considerations

6.1. Processing Non-ASCII Strings

Existing DNS servers do not know the IDNA rules for handling non-ASCII forms of IDNs, and therefore need to be shielded from them. All existing channels through which names can enter a DNS server database (for example, master files (as described in RFC 1034) and DNS update messages [RFC2136]) are IDN-unaware because they predate IDNA. Other sections of this document provide the needed shielding by ensuring that internationalized domain names entering DNS server databases through such channels have already been converted to their equivalent ASCII A-label forms.

Because of the design of the algorithms in Section 4 and Section 5 (a domain name containing only ASCII codepoints can not be converted to an A-label), there can not be more than one A-label form for each U-label.

The current update to the definition of the DNS protocol [RFC2181] explicitly allows domain labels to contain octets beyond the ASCII range (0000..007F), and this document does not change that. Note, however, that there is no defined interpretation of octets 0080..00FF as characters. If labels containing these octets are returned to applications, unpredictable behavior could result. The A-label form, which cannot contain those characters, is the only standard representation for internationalized labels in the current DNS protocol.

6.2. DNSSEC Authentication of IDN Domain Names

DNS Security [RFC2535] is a method for supplying cryptographic verification information along with DNS messages. Public Key Cryptography is used in conjunction with digital signatures to provide a means for a requester of domain information to authenticate the source of the data. This ensures that it can be traced back to a trusted source, either directly or via a chain of trust linking the source of the information to the top of the DNS hierarchy.

IDNA specifies that all internationalized domain names served by DNS servers that cannot be represented directly in ASCII must use the A-label form. Conversion to A-labels must be performed prior to a zone being signed by the private key for that zone. Because of this ordering, it is important to recognize that DNSSEC authenticates a domain name containing A-labels or conventional LDH-labels, not U-labels. In the presence of DNSSEC, no form of a zone file or query response that contains a U-label may be signed or the signature validated.

One consequence of this for sites deploying IDNA in the presence of DNSSEC is that any special purpose proxies or forwarders used to transform user input into IDNs must be earlier in the resolution flow than DNSSEC authenticating nameservers for DNSSEC to work.

6.3. Root Server Considerations

IDNs in A-label form will generally be somewhat longer than current domain names, so the bandwidth needed by the root servers is likely to go up by a small amount. Also, queries and responses for IDNs will probably be somewhat longer than typical queries today, so more queries and responses may be forced to go to TCP instead of UDP.

7. Security Considerations

The general security principles and issues for IDNA appear in [IDNA2008-Rationale]. The comments below are specific to this pair of protocols, but should be read in the context of that material and the definitions and specifications, identified there, on which this one depends.

This memo describes procedures for registering and looking up labels that are not valid according to the base DNS specifications (STD13 [RFC1034] [RFC1035] and Host Requirements [RFC1123]) because they contain non-ASCII characters. These procedures depend on the use of a special ASCII-compatable encoded form that contains only characters permitted in host names by those earlier specifications. The

encoding is specified in [RFC3492]. No security issues such as string length increases or new allowed values are introduced by the encoding process or the use of these encoded values, apart from those introduced by the ACE encoding itself.

Domain names (or portions of them) are sometimes compared against a set of privileged or anti-privileged domains. In such situations it is especially important that the comparisons be done properly, as specified in requirement 2 of Section 3.1. For labels already in ASCII form (i.e., are LDH-labels or A-labels), the proper comparison reduces to the same case-insensitive ASCII comparison that has always been used for ASCII labels.

The introduction of IDNA means that any existing labels that start with the ACE prefix would be construed as A-labels, at least until they failed one of the relevant tests, whether or not that was the intent of the zone administrator or registrant. There is no evidence that this has caused any practical problems since RFC 3490 was adopted, but the risk still exists in principle.

8. IANA Considerations

IANA actions for this version of IDNA are specified in [IDNA2008-Rationale].

9. Change Log

[[anchor23: RFC Editor: Please remove this section.]]

9.1. Version -00 of draft-klensin-idnabis-protocol

Version -00 of this draft was produced in November 2007 by moving text from draft-klensin-idnabis-issues and by copy considerable text from RFC 3490. The result was then extensively edited.

9.2. Versions -01 and -02 of draft-klensin-idnabis-protocol

These versions reflected a number of editorial changes, some of them significant, and alignment of terminology with draft-faltstrom-idnabis-tables.

9.3. Version -03 of draft-klensin-idnabis-protocol

- Abstract rewritten to bring its length within RFC Editor guidelines.

- Corrections and revisions in response to extensive comments by Mark Davis and others.

- Small modifications to several operations, including moving the Normalization steps to a different place in the sequence.

- Many editorial changes.

9.4. Version -04 of draft-klensin-idnabis-protocol

- Revised terminology and removed the MAYBE category as a consequence of design discussions on 30 January 2003 and followup conversations. Also restructured the various operations to treat CONTEXTUAL RULE REQUIRED as a validation step (paralleling bidi), rather than a category. Those changes required changes elsewhere in the document for consistency.

- Changed the requirements for normalization, making this a requirement on the calling application rather than an action of this protocol. This is consistent with the general "mappings belong somewhere else" principle.

- Updated references.

- More editorial work, some independent of the changes, described immediately above.

9.5. Version -00 of draft-ietf-idnabis-protocol

- Clarified actions to be taken if an A-label is supplied as input.

- Moved the contextual rules appendix into this document from draft-klensin-idnabis-issues and made an initial attempt at defining the actual rules. Synchronized the list of characters in that appendix with tables-01.

- Added an explicit discussion of A-label input.

- Inserted a test for double-hyphen here.

10. Contributors

While the listed editor held the pen, the original versions of this document represent the joint work and conclusions of an ad hoc design team consisting of the editor and, in alphabetic order, Harald Alvestrand, Tina Dam, Patrik Faltstrom, and Cary Karp. This document draws significantly on the original version of IDNA [RFC3490] both

conceptually and for specific text. This second-generation version would not have been possible without the work that went into that first version and its authors, Patrik Faltstrom, Paul Hoffman, and Adam Costello. While Faltstrom was actively involved in the creation of this version, Hoffman and Costello were not and should not be held responsible for any errors or omissions.

11. Acknowledgements

This revision to IDNA would have been impossible without the accumulated experience since RFC 3490 was published and resulting comments and complaints of many people in the IETF, ICANN, and other communities, too many people to list here. Nor would it have been possible without RFC 3490 itself and the efforts of the Working Group that defined it. Those people whose contributions are acknowledged in RFC 3490, [RFC4690], and [IDNA2008-Rationale] were particularly important.

12. References

12.1. Normative References

[IDNA2008-BIDI]
Alvestrand, H. and C. Karp, "An updated IDNA criterion for right-to-left scripts", January 2008, <http://www.ietf.org/internet-drafts/draft-alvestrand-idna-bidi-03.txt>.

[IDNA2008-Rationale]
Klensin, J., Ed., "Internationalizing Domain Names for Applications (IDNA): Issues, Explanation, and Rationale", February 2008, <http://www.ietf.org/internet-drafts/draft-klensin-idnabis-issues-06.txt>.

[IDNA2008-Tables]
Faltstrom, P., "The Unicode Codepoints and IDN", February 2008, <http://stupid.domain.name/idnabis/draft-faltstrom-idnabis-tables-04.txt>.

A version of this document, is available in HTML format at http://stupid.domain.name/idnabis/draft-faltstrom-idnabis-tables-04.html

[RFC1034] Mockapetris, P., "Domain names - concepts and facilities", STD 13, RFC 1034, November 1987.

[RFC1035] Mockapetris, P., "Domain names - implementation and specification", STD 13, RFC 1035, November 1987.

[RFC1123] Braden, R., "Requirements for Internet Hosts - Application and Support", STD 3, RFC 1123, October 1989.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC3492] Costello, A., "Punycode: A Bootstring encoding of Unicode for Internationalized Domain Names in Applications (IDNA)", RFC 3492, March 2003.

[Unicode-PropertyValueAliases]
The Unicode Consortium, "Unicode Character Database: PropertyValueAliases", March 2008, <http://www.unicode.org/Public/UNIDATA/PropertyValueAliases.txt>.

[Unicode-RegEx]
The Unicode Consortium, "Unicode Technical Standard #18: Unicode Regular Expressions", May 2005, <http://www.unicode.org/reports/tr18/>.

[Unicode-Scripts]
The Unicode Consortium, "Unicode Standard Annex #24: Unicode Script Property", February 2008, <http://www.unicode.org/reports/tr24/>.

[Unicode-UAX15]
The Unicode Consortium, "Unicode Standard Annex #15: Unicode Normalization Forms", 2006, <http://www.unicode.org/reports/tr15/>.

12.2. Informative References

[ASCII] American National Standards Institute (formerly United States of America Standards Institute), "USA Code for Information Interchange", ANSI X3.4-1968, 1968.

ANSI X3.4-1968 has been replaced by newer versions with slight modifications, but the 1968 version remains definitive for the Internet.

[RFC2136] Vixie, P., Thomson, S., Rekhter, Y., and J. Bound, "Dynamic Updates in the Domain Name System (DNS UPDATE)", RFC 2136, April 1997.

[RFC2181] Elz, R. and R. Bush, "Clarifications to the DNS

Specification", RFC 2181, July 1997.

[RFC2535] Eastlake, D., "Domain Name System Security Extensions", RFC 2535, March 1999.

[RFC3490] Faltstrom, P., Hoffman, P., and A. Costello, "Internationalizing Domain Names in Applications (IDNA)", RFC 3490, March 2003.

[RFC3986] Berners-Lee, T., Fielding, R., and L. Masinter, "Uniform Resource Identifier (URI): Generic Syntax", STD 66, RFC 3986, January 2005.

[RFC3987] Duerst, M. and M. Suignard, "Internationalized Resource Identifiers (IRIs)", RFC 3987, January 2005.

[RFC4690] Klensin, J., Faltstrom, P., Karp, C., and IAB, "Review and Recommendations for Internationalized Domain Names (IDNs)", RFC 4690, September 2006.

[RFC4952] Klensin, J. and Y. Ko, "Overview and Framework for Internationalized Email", RFC 4952, July 2007.

[Unicode] The Unicode Consortium, "The Unicode Standard, Version 5.0", 2007.

Boston, MA, USA: Addison-Wesley. ISBN 0-321-48091-0

Appendix A. The Contextual Rules Registry

[[anchor33: Note in Draft: the WG needs to figure out whether this table stays as part of this document, is moved to a separate one, or is incorporated into "tables". Regardless of where they are placed, the WG will still need to review the specific content of the rules. In this version of the document, the table remains something of a illustrative placeholder, not a final specification.]]

As discussed in the IANA Considerations section of [IDNA2008-Rationale], a registry of rules that define the contexts in which particular PROTOCOL-VALID characters, characters associated with a requirement for Contextual Information, are permitted. These rules are expressed as tests on the label in which the characters appear (all, or any part of, the label may be tested).

For each character specified as requiring a contextual rule, a rule MAY be established with the following data elements:

1. The code point associated with the character.

2. The name of the character.

3. An indication as to whether the code point requires the rule be processed at lookup time (this indication is equivalent to the difference between "CONTEXTJ" and "CONTEXTO" in the tables document [IDNA2008-Tables]).

4. A prose description of the contextual rule.

5. A description of the contextual rule using Unicode Regular Expression notation [Unicode-RegEx]. Only a Level 1 implementation is needed for the expressions below, which also make reference to the Unicode Script definition [Unicode-Scripts] and the Unicode Property Value Aliases list [Unicode-PropertyValueAliases]. Note that in these regular expressions, the label is taken to be an entire line, i.e., "^" refers to the beginning of the label and "$" refers to the end of the label.

   These regular expressions are used as tests. The contextual requirement is met if there is a match for the regular expression and not met if there is no match.

   [[anchor34: Patrik and I (JcK) would like to find a way to state these rules that does not require the reader and implementer to understand what we believe to be a fairly exotic element of the Unicode specification. Suggestions welcome.]]

6. An optional comment preceded by "#"

Should there be any conflict between the two statements of a rule, the regular expression form MUST be considered normative until the registry can be corrected.

The rules for the characters listed in the Tables document as exception cases or Join_Controls and for which rules are being defined at this time appear below.

[[anchor35: Note in draft: This table is not complete and the rule entries below are temporarily only examples.]]

```
002D; HYPHEN-MINUS; F;
   Must not appear at the beginning or end of a label;
   Regular expression:
   [^^]\u002D|\u00SD[^$] ;
   # Note that a prohibition on having two hyphens as the third and
```

fourth characters of anything but a valid A-label appears in the specification.

200C; ZERO WIDTH NON-JOINER; T;
Between two characters from the same script only. The script must be one in which the use of this character causes significant visual transformation of one or both of the adjacent characters;
Regular expression:
[\p(Script:Deva)\p(Script:Tamil)]\u200C[\p(Script:Deva)\p(Script:Tamil)] ;
[[anchor36: That script list is _not_ complete and, in particular, more Indic scripts certainly need to be listed. It also does not correctly express the "same script" restriction mentioned in the prose, since it only tests adjacent characters. Whether this character is required for Arabic script, and with what restrictions if it is, is under discussion in the WG and in other forums. It is clear that a Unicode derived property for script groups that would permit testing, e.g., "Indic Script", would be very helpful here.]]

200D; ZERO WIDTH JOINER; T;
Between two characters from the same script only. The script must be one in which the use of this character causes significant visual transformation of one or both of the adjacent characters;
Regular expression:
[\p(Script:Deva)\p(Script:Tamil)]+
\u200D[\p(Script:Deva)\p(Script:Tamil)]+ ;
[[anchor37: That script list is _not_ complete and, in particular, more Indic scripts certainly need to be listed. It also does not correctly express the "same script" restriction mentioned in the prose, since it only tests adjacent characters. Whether this character is required for Arabic script, and with what restrictions if it is, is under discussion in the WG and in other forums.]]

00B7; MIDDLE DOT; F;
Between two 'l' (U+006C) characters only, used to permit the Catalan character ela geminada to be expressed;
Regular expression:
\u006C\u00B7\u006c ;

0375; GREEK LOWER NUMERAL SIGN (KERAIA); F;
Greek script only. Might be further restricted to specific following characters;
Regular expression:
\0375\(Script:Greek) ;

```
   02B9; MODIFIER LETTER PRIME; F;;;
      # Permitted only in contexts in which GREEK LOWER NUMERAL SIGN,
      U+0375, is permitted.  GREEK NUMERAL SIGN, U+0374, and the Lower
      Numeral Sign (U+0375) are indicators for numeric use of letters in
      older Greek writing systems.  U+02B9 is relevant because
      normalization maps U+0374 into it.;
      Regular expression:
      \(Script:Greek)\02B9\(Script:Greek) ;
      [[anchor38: The test is that the adjacent characters be in the
      Greek script.  It is not clear whether this is sufficient.  The
      requirement for a preceding Greek letter may not be necessary.
      More input needed.]]

   0483; COMBINING CYRILLIC TITLO; F;
      Cyrillic script only.  Might be further restricted to permit only
      a preceding list of characters.
      Regular expression:
      \p(Script:Cyrillic)\u0483 ;

   05F3; HEBREW PUNCTUATION GERESH; F;
      The script of the preceding character and the subsequent
      character, if any, MUST be Hebrew;
      Regular expression:
      \p(Script:Hebrew)\u05F3\p(Script:Hebrew)? ;

   05F4; HEBREW PUNCTUATION GERSHAYIM; F
      The script of the preceding character and the subsequent
      character, if any, MUST be Hebrew;
      Regular expression:
      \p(Script:Hebrew)\u05F3\p(Script:Hebrew)? ;

   3005; IDEOGRAPHIC ITERATION MARK; F;
      MUST NOT be at the beginning of the label, and the previous
      character MUST be in Han Script;
      Regular expression:
      \p(Script:Hani)\u30FB ;

   303B; VERTICAL IDEOGRAPHIC ITERATION MARK; F;
      MUST NOT be at the beginning of the label, and the previous
      character MUST be in Han Script;
      Regular expression:
      \p(Script:Hani)\u303B ;

   30FB; KATAKANA MIDDLE DOT; F;
      Adjacent characters MUST be Katakana;
      Regular expression:
      \p(Script:Kana)\u30FB\p(Script:Kana) ;
```

While the information above is to be used to initialize the registry, IANA should treat the table format in this Appendix simply as an initial, tentative, suggestion. Subject to review and comment from the IESG and any Expert Reviewers, IANA is responsible for, and should develop, a format for that registry, or a copy of it maintained in parallel, that is convenient for retrieval and machine processing and publish the location of that version.

Author's Address

John C Klensin (editor)
1770 Massachusetts Ave, Ste 322
Cambridge, MA 02140
USA

Phone: +1 617 245 1457
Fax:
Email: john+ietf@jck.com
URI:

Full Copyright Statement

Intellectual Property