                 The Use of Non-ASCII Characters in RFCs
                        draft-flanagan-nonascii-05

Abstract

   In order to support the internationalization of protocols and a
more
   diverse Internet community, the RFC Series must evolve to allow for
   the use of non-ASCII characters in RFCs. While English remains the
   required language of the Series, the encoding of future RFCs will
be
   in UTF-8, allowing for a broader range of characters than typically
   used in the English language. This document describes the RFC
Editor
   requirements and guidance regarding the use of non-ASCII characters
   in RFCs.

   This document updates RFC 7322.

Status of This Memo

This Internet-Draft will expire on February 20, 2016.

Table of Contents

1. Introduction

   For much of the history of the RFC Series, the character encoding
   used for RFCs has been ASCII [ANSI.X3-4.1986]. This was a sensible
   choice at the time: the language of the Series has always been
   English, a language that primarily uses ASCII-encoded characters
   (ignoring for a moment words borrowed from more richly decorated
   alphabets); and, ASCII is the "lowest common denominator" for
   character encoding, making cross-platform viewing trivial.

   There are limits to ASCII, however, that hinder its continued use
as
   the exclusive character encoding for the Series. The increasing
need
   for easily readable, internationalized content suggests it is time
to
   allow non-ASCII characters in RFCs where necessary. To support
this
   move away from ASCII, RFCs will switch to supporting UTF-8 as the

default character encoding and allow support for a broad range of
Unicode character support. [UnicodeCurrent] Note that the RFC
Editor may reject any codepoint that does not render adequately in
enough formats or on in enough rendering engines using the current
tooling.

Given the continuing goal of maximum readability across platforms,
the use of non-ASCII characters should be limited in a document to
only where necessary within the text. This document describes the
rules under which non-ASCII characters may be used in an RFC.
These
rules will be applied as the necessary changes are made to
submission
checking and editorial tools.

This document updates the RFC Style Guide [RFC7322].

The details described in this document are expected to change based
on experience gained in implementing the RFC production center's
toolset. Revised documents will be published capturing those
changes
as the toolset is completed. Other implementers must not expect
those changes to remain backwards-compatible with the details
described this document.

2. Basic requirements

Two fundamental requirements inform the guidance and examples
provided in this document. They are:

o Searches against RFC indexes and database tables need to return
   expected results and support appropriate Unicode string matching
   behaviors;

o RFCs must be able to display correctly across a wide range of
   readers and browsers. People whose system does not have the
fonts
   needed to display a particular RFC need to be able to read the
   various publication formats and the XML correctly in order to

understand and implement the information described in the
      document.

3. Rules for the use of non-ASCII characters

   This section describes the guidelines for the use of non-ASCII
   characters in the header, body, and reference sections of an RFC.
If
   the RFC Editor identifies areas where the use of non-ASCII
characters
   negatively impacts the readability of the text, they will request
   alternate text.

   The RFC Editor may, in cases of entire words represented in non-
ASCII
   characters, ask for a set of reviewers to verify the meaning,
   spelling, characters, and grammar of the text.

3.1. General usage throughout a document

   Where the use of non-ASCII characters is purely as part of an
example
   and not otherwise required for correct protocol operation, escaping
   the non-ASCII character is not required. Note, however, that as
the
   language of the RFC Series is English, the use of non-ASCII
   characters is based on the spelling of words commonly used in the
   English language following the guidance in the Merriam-Webster
   dictionary [MerrWeb].

   The RFC Editor will use the primary spelling listed in that
   dictionary by default.

   Example of non-ASCII characters that do not require escaping
   [RFC4475]:

This particular response contains unreserved and non-ascii

UTF-8 characters.
This response is well formed. A parser must accept this message.
Message Details : unreason
SIP/2.0 200 = 2**3 * 5**2 но сто девяносто девять – простое
Via: SIP/2.0/UDP 192.0.2.198;branch=z9hG4bK1324923
Call-ID: unreason.1234ksdfak3j2erwedfsASdf
CSeq: 35 INVITE
From: sip:user@example.com;tag=11141343
To: sip:user@example.edu;tag=2229 Content-Length: 154
Content-Type: application/sdp

3.2. Authors, Contributors, and Acknowledgments

   Person names may appear in several places within an RFC. In all
   cases, valid Unicode is required. For names that include non-ASCII
   characters, an author-provided, ASCII-only identifier is required
to
   assist in search and indexing of the document.

   Example for the header:

   Example for the Acknowledgements:

OLD: The following people contributed significant text to early
versions of this draft: Patrik Faltstrom, William Chan, and Fred
Baker.

PROPOSED/NEW: The following people contributed significant text to
early versions of this draft: Patrik Fältström (Patrik Faltstrom),
陈智昌 (William Chan), and Fred Baker.

3.3. Company Names

   Company names may appear in several places within an RFC. The
rules
   for company names follow similar guidance to that of person names.
   Valid Unicode is required. For company names that include non-
ASCII
   characters, an ASCII-only identifier is required to assist in
search
   and indexing of the document.

3.4. Body of the document

   When the mention of non-ASCII characters is required for correct
   protocol operation and understanding, the characters' Unicode
   character name or code point MUST be included in the text.

   o Non-ASCII characters will require identifying the Unicode code
      point.

   o Use of the actual UTF-8 character (e.g., Δ) is encouraged so
      that a reader can more easily see what the character is, if
their
      device can render the text.

   o The use of the Unicode character names like "INCREMENT" in
      addition to the use of Unicode code points is also encouraged.
      When used, Unicode character names should be in all capital
      letters.

   Examples:

```
   OLD [RFC7564]:

   However, the problem is made more serious by introducing the full
   range of Unicode code points into protocol strings.  For example,
   the characters U+13DA U+13A2 U+13B5 U+13AC U+13A2 U+13AC U+13D2
from
   the Cherokee block look similar to the ASCII characters  "STPETER"
as
   they might appear when presented using a "creative" font family.

   NEW/ALLOWED:

However, the problem is made more serious by introducing the full
range of Unicode code points into protocol strings.  For example,
the characters U+13DA U+13A2 U+13B5 U+13AC U+13A2 U+13AC U+13D2
(ST𝖯ETER) from the Cherokee block look similar to the ASCII
characters "STPETER" as they might appear when presented using a
"creative" font family.

   ALSO ACCEPTABLE:

However, the problem is made more serious by introducing the full
range of Unicode code points into protocol strings.  For example,
the characters "ST𝖯ETER" (U+13DA U+13A2 U+13B5 U+13AC U+13A2
U+13AC U+13D2) from the Cherokee block look similar to the ASCII
characters "STPETER" as they might appear when presented using a
"creative" font family.

   Example of proper identification of Unicode characters in an RFC:

   Acceptable:

   o  Temperature changes in the Temperature Control Protocol are
      indicated by the U+2206 character.

   Preferred:

   1.  Temperature changes in the Temperature Control Protocol are
       indicated by the U+2206 character ("∆").

   2.  Temperature changes in the Temperature Control Protocol are
       indicated by the U+2206 character (INCREMENT).

   3.  Temperature changes in the Temperature Control Protocol are
       indicated by the U+2206 character ("∆", INCREMENT).

   4.  Temperature changes in the Temperature Control Protocol are
       indicated by the U+2206 character (INCREMENT, "∆").
```

     5. Temperature changes in the Temperature Control Protocol are
         indicated by the [Delta] character "Δ" (U+2206).

     6. Temperature changes in the Temperature Control Protocol are
         indicated by the character "Δ" (INCREMENT, U+2206).

     Which option of (1), (2), (3), (4), (5), or (6) is preferred may
     depend on context and the specific character(s) in question. All
are
     acceptable within an RFC. BCP 137, "ASCII Escaping of Unicode
     Character" describes the pros and cons of different options for
     identifying Unicode characters in an ASCII document BCP137
[RFC5137].

3.5. Tables

     Tables follow the same rules for identifiers and characters as in
     "Section 3.4. Body of the document". If it is sensible (i.e.,
more
     understandable for a reader) for a given document to have two
tables
     -- one including the identifiers and non-ASCII characters and a
     second with just the non-ASCII characters -- that will be allowed
on
     a case-by-case basis.

     Original text from "Preparation, Enforcement, and Comparison of
     Internationalized Strings Representing Usernames and Passwords"
     [RFC7613].

     Table 3: A sample of legal passwords

        +------------------------------------
+-----------------------------+
        | # | Password | Notes
|
        +------------------------------------

```
       +-------------------------------+
       | 12| <correct horse battery staple> | ASCII space is allowed
       |
          +---------------------------------
       +-------------------------------+
       | 13| <Correct Horse Battery Staple> | Different from example 12
       |
          +---------------------------------
       +-------------------------------+
       | 14| <&#x3C0;&#xDF;&#xE5;>          | Non-ASCII letters are OK
       |
       |   |                                | (e.g., GREEK SMALL LETTER
       |
       |   |                                | PI, U+03C0)
       |
          +---------------------------------
       +-------------------------------+
       | 15| <Jack of &#x2666;s>            | Symbols are OK (e.g., BLACK
       |
       |   |                                | DIAMOND SUIT, U+2666)
       |
          +---------------------------------
       +-------------------------------+
       | 16| <foo&#x1680;bar>               | OGHAM SPACE MARK, U+1680, is
       |
       |   |                                | mapped to U+0020 and thus
       |
       |   |                                | the full string is mapped to
       |
       |   |                                | <foo bar>
       |
          +---------------------------------
       +-------------------------------+

       Preferred text:
```

```
    Table 3: A sample of legal passwords

    +-----------------------------------+-------------------------------+
    | # | Password                      | Notes                         |
```

```
+-------------------------------+-------------------------------+
| 12| <correct horse battery staple> | ASCII space is allowed        |
+-------------------------------+-------------------------------+
| 13| <Correct Horse Battery Staple> | Different from example 12     |
+-------------------------------+-------------------------------+
| 14| <πßഌ>                     | Non-ASCII letters are OK      |
|   |                           | (e.g., GREEK SMALL LETTER     |
|   |                           | PI, U+03C0; LATIN SMALL       |
|   |                           | LETTER SHARP S, U+00DF; THAI  |
|   |                           | DIGIT SEVEN, U+0E57)          |
+-------------------------------+-------------------------------+
| 15| <Jack of ♦s>              | Symbols are OK (e.g., BLACK   |
|   |                           | DIAMOND SUIT, U+2666)         |
+-------------------------------+-------------------------------+
| 16| <foo-bar>                 | OGHAM SPACE MARK, U+1680, is  |
|   |                           | mapped to U+0020 and thus     |
|   |                           | the full string is mapped to  |
|   |                           | <foo bar>                     |
+-------------------------------+-------------------------------+
```

3.6. Code components

   The RFC Editor encourages the use of the U+ notation except within
a
   code component where you must follow the rules of the programming
   language in which you are writing the code.

3.7. Bibliographic text

   The reference entry must be in English; whatever subfields are
   present must be available in ASCII-encoded characters. As long as
   good sense is used, the reference entry may also include non-ASCII
   characters at the author's discretion and as provided by the
author.
   The RFC Editor may request a review of the non-ASCII reference
entry.
   This applies to both normative and informative references.

   Example:

[GOST3410] "Information technology. Cryptographic data security.
          Signature and verification processes of [electronic]
          digital signature.", GOST R 34.10-2001, Gosudarstvennyi
          Standard of Russian Federation, Government Committee of
          Russia for Standards, 2001. (In Russian)

Allowable addition to the above citation:
          "Информационная технология. Криптографическая защита
          информации. Процессы формирования и проверки
          электронной цифровой подписи", GOST R 34.10-2001,
          Государственный стандарт Российской Федерации, 2001.

3.8. Keywords and Citation Tags

   Keywords and citation tags must be ASCII only.

3.9. Address Information

   The purpose of providing address information, either postal or
   e-mail, is to assist readers of an RFC to contact the author or
   authors. Authors may include the official postal address as
   recognized by their company or local postal service without
   additional non-ASCII character escapes. If the email address
   includes non-ASCII characters and is a valid email address at the
   time of publication, non-ASCII character escapes are not required.

4. Normalization Forms

   Authors should not expect normalization forms to be preserved. If
a
   particular normalization form is expected, note that in the text of
   the RFC.

5. XML Markup

   As described above, use of non-ASCII characters in areas such as
   email, company name, addresses, and name is allowed. In order to
   make it easier for code to identify the appropriate ASCII
   alternatives, authors must include an "ascii" attribute to their
XML
   markup. See [I-D.hoffman-xml2rfc] for more detail.

6. IANA Considerations

   This document makes no request of IANA.

   Note to RFC Editor: this section may be removed on publication as an
   RFC.

7. Internationalization Considerations

   The ability to use non-ASCII characters in RFCs in a clear and
   consistent manner will improve the ability to describe
   internationalized protocols and will recognize the diversity of
   authors. However, the goal of readability will override the use of
   non-ASCII characters within the text.

8. Security Considerations

   Valid Unicode that matches the expected text must be verified in
   order to preserve expected behavior and protocol information.

9. Change log - to be removed by the RFC Editor

9.1. -04 to -05

   Keywords: expanded section to include citation tags.

   Internationalization considerations: reiterated that the use of non-
   ASCII characters is not automatically guaranteed.

9.2. -04 to -05

   Introduction: added statement regarding document subject to change.

   Tables: added example.

   Code: removed placeholder for example.

9.3. -02 to -04

Introduction and Abstract: change to be clearer about what/why non-ASCII characters are being allowed.

XML Markup: section added.

10. References

[ANSI.X3-4.1986]
          American National Standards Institute, "Coded Character
          Set - 7-bit American Standard Code for Information
          Interchange", ANSI X3.4, 1986.

[I-D.hoffman-xml2rfc]
          Hoffman, P., "The 'XML2RFC' version 3 Vocabulary",
draft-
          hoffman-xml2rfc-21 (work in progress), July 2015.

[MerrWeb] Merriam-Webster,Inc., "Merriam-Webster's Collegiate
          Dictionary, 11th Edition", 2009.

[RFC3550] Schulzrinne, H., Casner, S., Frederick, R., and V.
          Jacobson, "RTP: A Transport Protocol for Real-Time
          Applications", STD 64, RFC 3550, DOI 10.17487/RFC3550,
          July 2003, <http://www.rfc-editor.org/info/rfc3550>.

[RFC4475] Sparks, R., Ed., Hawrylyshen, A., Johnston, A.,
Rosenberg,
          J., and H. Schulzrinne, "Session Initiation Protocol
(SIP)
          Torture Test Messages", RFC 4475, DOI 10.17487/RFC4475,
          May 2006, <http://www.rfc-editor.org/info/rfc4475>.

[RFC5137] Klensin, J., "ASCII Escaping of Unicode Characters",
          BCP 137, RFC 5137, DOI 10.17487/RFC5137, February 2008,
          <http://www.rfc-editor.org/info/rfc5137>.

   [RFC6949] Flanagan, H. and N. Brownlee, "RFC Series Format
              Requirements and Future Development", RFC 6949,
              DOI 10.17487/RFC6949, May 2013,
              <http://www.rfc-editor.org/info/rfc6949>.

   [RFC7322] Flanagan, H. and S. Ginoza, "RFC Style Guide", RFC 7322,
              DOI 10.17487/RFC7322, September 2014,
              <http://www.rfc-editor.org/info/rfc7322>.

   [RFC7564] Saint-Andre, P. and M. Blanchet, "PRECIS Framework:
              Preparation, Enforcement, and Comparison of
              Internationalized Strings in Application Protocols",
              RFC 7564, DOI 10.17487/RFC7564, May 2015,
              <http://www.rfc-editor.org/info/rfc7564>.

   [RFC7613] Saint-Andre, P. and A. Melnikov, "Preparation,
              Enforcement, and Comparison of Internationalized Strings
              Representing Usernames and Passwords", RFC 7613,
              DOI 10.17487/RFC7613, August 2015,
              <http://www.rfc-editor.org/info/rfc7613>.

   [UnicodeCurrent]
              The Unicode Consortium, "The Unicode Standard",
              2014-present, <http://www.unicode.org/versions/latest/>.

Appendix A. Acknowledgements

   With many thanks to the members of the IAB i18n program and the RFC
   Format Design Team.

Author's Address

   Heather Flanagan (editor)
   RFC Editor

   Email: rse@rfc-editor.org
   URI: http://orcid.org/0000-0002-2647-2220