Network Working Group
Internet-Draft
Updates: 2223 (if approved)
Intended status: Informational
Expires: September 30, 2010

P. Hoffman
VPN Consortium
T. Bray
Sun Microsystems

March 29, 2010

Using non-ASCII Characters in RFCs
draft-hoffman-utf8-rfcs-06

Abstract

This document specifies a change to the IETF process in which Internet Drafts and RFCs are allowed to contain non-ASCII characters. The proposed change is to change the encoding of Internet Drafts and RFCs to UTF-8 when non-ASCII characters are needed.

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt.

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html.

This Internet-Draft will expire on September 30, 2010.

1. Introduction

The purpose of this document is to specify a way for the IETF to use non-ASCII characters in Internet Drafts and RFCs.

Various guideline documents in the IETF, notably [RFC2223], specify that RFCs must use only the US-ASCII character set. This restriction has historically caused problems, notably:

- Names and addresses of authors of IETF documents are misspelled
- Names and document titles in references are misspelled
- Protocol examples that include non-ASCII characters cannot be included straightforwardly

The first two issues cause real problems for people searching for RFCs for particular authors or references that contain non-ASCII characters. For many languages that use Latin characters outside the ASCII range, there are no absolute mappings between those non-ASCII characters and ASCII equivalents. A common example is that "u-with-umlaut" (U+00FC) may be mapped to "u" or to "ue"; many other mapping difficulties exist.

The third issue reduces the effectiveness of IETF specifications; implementors of protocols which carry textual payloads often experience difficulty in achieving interoperability related to the use of character sets from around the world. Specifications which can provide concrete examples of such protocol scenarios will be of significant benefit to these implementors.

Now that UTF-8 [RFC3629] is nearly universally available in text-editing and display systems, the IETF can eliminate these problems by allowing RFCs to use UTF-8. As a reminder, UTF-8 is fully and thoroughly upwards compatible to US-ASCII.

This document uses example characters as specified in [RFC5137]. Had the recommendations from this document already been implemented, this alternate representation would, of course, not be necessary.

It is important to note that this document does not use RFC 2119 language (MUST, SHOULD, and so on). Instead, it lists practices that the IETF should consider. If the ideas in this document are adopted, the final list of rules for using UTF-8 in Internet Drafts and RFCs would be published by the IETF Secretariat and the RFC Editor. The authors are open to changing this and using 2119-style language if the community prefers it.

2. Use of UTF-8 in Internet Drafts and RFCs

Upon publication of this document as an RFC, new RFCs and Internet Drafts will be considered to be encoded in UTF-8 if they contain any non-ASCII characters; otherwise, they will continue to be considered encoded in US-ASCII. The IETF Secretariat and RFC Editor need to change their processes to publish documents that are valid UTF-8.

2.1. Limits On the Locations In Which Non-ASCII Text May Be Used

It is suggested that the IETF Secretariat and RFC Editor limit non-ASCII characters to the following:

- Names and addresses of authors, used at the top of RFCs and in Author Contact sections
- Names and document titles used in References sections
- Quotations where the original contains non-ASCII characters
- Protocol examples that include non-ASCII characters, for example in Internationalized Domain Names (IDNs), Internationalized Resource Identifiers (IRIs), and Internationalized Email Addresses (IEA).

Using non-ASCII characters in areas other than those listed above is prohibited. In specific, using "curly quotes", m-dashes, and other punctuation that appear in normal publishing is not allowed under these guidelines. This limitation is to help those people who are reading Internet Drafts and RFCs on systems that do not render UTF-8 legibly.

2.2. Allowable Character Repertoire

UTF-8 is an encoding of the Unicode Character Set and can be used to encode any of its numeric codepoints, from U+0000 to U+10FFFF inclusive. Specifications using UTF-8 must not use the following codepoint ranges:

- The "ASCII control characters" in the ranges U+0000 to U+0008, U+000B, and U+000D to U+001F. Also, the "C1 control characters" in the ranges U+0080 to U+009F. These lack either visual representations, interoperable semantics, or both.
- The Surrogate-block range U+D800 to U+DFFF. These codepoints do

not identify characters, but exist to support the UTF-16 encoding.

- The ZERO WIDTH NO-BREAK SPACE U+FEFF and its mirror image U+FFFE.

- The Private-Use-Area ranges, U+E000 to U+F8FF, U+F0000 to U+FFFFD, and U+100000 to U+10FFFD.

Internet Drafts and RFCs should not contain Unicode codepoints which are "Compatibility Characters", that is, those whose properties include a compatibility decomposition. Note that such characters occur rarely and detecting them requires run-time access to the Unicode character database, which may not be practical in some situations.

[[Need to add additional types of characters that should not be allowed: unassigned characters, other control characters, ones that are really formatting characters, and maybe others. This needs some wording, given that the lists of these change over time.]]

2.3. Normalization

Due to the way that Unicode uses combining characters, there are sometimes multiple codepoint sequences that denote what, to a human, is the same character. For example, the character "lowercase-a-with-acute" can be spelled in two ways: as a single character (U+00E1) or as two characters (U+0061 followed by U+0301). This can present problems in searching and rendering.

The process of standardizing on one of these possibilities is referred to as "normalization" and several "normalization forms" are defined by the Unicode Consortium. All UTF-8 text appearing in RFCs (but not necessarily Internet Drafts) ought to be normalized using Normalization Form C [[reference needed, should be the version of Unicode when this is finalized]].

2.4. Author and Employer Names

Authors can choose how to spell their names and the names of their employers in the various parts of Internet Drafts they are writing. The spelling at the top of the first page of the document needs to match the spelling in the "Authors' Addresses" section near the end of the document, but the latter can have alternate spellings to help those searching documents by name. Postal information listed in the "Authors' Addresses" section can also use non-ASCII.

For example, assume that an author whose name is <U+6653><U+4E1C> F<U+00E4>ltstr<U+00F6>m has a preferred all-ASCII spelling of Xiaodong Faltstrom. One expected allowed methods for spelling his name would be:

```
Network Working Group                                  X. Faltstrom
Internet-Draft                                            ExampleCo
. . .
Author's Address
```

```
    Xiaodong Faltstrom (<U+6653><U+4E1C> F<U+00E4>ltstr<U+00F6>m)
    ExampleCo

    Email: xiaodong.faltstrom@example.com
```

Another expected allowed methods for spelling his name would be:

```
Network Working Group                   X. F<U+00E4>ltstr<U+00F6>m
Internet-Draft                                            ExampleCo
. . .
Author's Address

    <U+6653><U+4E1C> F<U+00E4>ltstr<U+00F6>m (Xiaodong Faltstrom)
    ExampleCo

    Email: xiaodong.faltstrom@example.com
```

3. Document Content

In order to assist text display software, any Internet Draft or RFC that contains non-ASCII characters should start with the byte order mark (BOM) U+FEFF. The UTF-8 byte order mark should not be included in any Internet Draft or RFC that does not contain non-ASCII characters. Detecting if an Internet Draft or RFC contains non-ASCII characters and being sure that such a document has a byte order mark can be done by the IETF's Internet Draft submission tool and the RFC Editor's publishing process.

RFCs are currently published with form-feed characters between pages. These marks work on some printers but not others. This proposed change does not affect any policy whether or not to use form-feed characters.

4. Security Considerations

A display program that expects only US-ASCII input may fail when it encounters octets outside the US-ASCII range of values. Such a failure may become a security issue. For example, the program may display incorrect results for the input. More seriously, the program may have an internal error that causes it to fail in a security-compromising fashion. Note that such a program is vulnerable to many attacks other than just showing IETF documents.

Someone could insert a UTF-8 host name in an RFC that has visually confusing characters. Another person could copy that host name out of the RFC and have it resolve to an unintended DNS name. This scenario seems quite far-fetched, given that tracking the RFC back to the author is trivial.

5. IAOC and IAB considerations

If this document is adopted by the IETF, it will be up to the IAOC and IAB to have the IETF Secretariat and the RFC Editor, respectively, implement it. The two bodies need to consider all of the suggested rules in this document, both the positive ones (such as allowing additional characters in some parts of Internet Drafts and RFCs) and the negative ones (such as disallowing particular characters from being used). The IAOC and IAB might want to publish proposed instructions to the IETF Secretariat and the RFC Editor and ask for community input on the specific instructions.

6. Informative References

[RFC2223] Postel, J. and J. Reynolds, "Instructions to RFC Authors", RFC 2223, October 1997.

[RFC3629] Yergeau, F., "UTF-8, a transformation format of ISO 10646", STD 63, RFC 3629, November 2003.

[RFC5137] Klensin, J., "ASCII Escaping of Unicode Characters", BCP 137, RFC 5137, February 2008.

Appendix A. Arguments Against Changing to UTF-8

Over more than a decade, the question of changing the encoding of RFCs to UTF-8 has come up repeatedly. Although many people wanted the change, various people had different reasons why they felt it was a bad idea. This appendix is a summary of those arguments and an explanation of why they are no longer as critical as they were long ago.

A.1. Difficulty in Displaying

Some text display systems only know how to display US-ASCII. Displaying an RFC that uses non-ASCII characters encoded in UTF-8 will cause those characters to be unreadable.

There are, of course, still such display systems, and there always will be. However, the number is dwindling as more software is improved to display non-ASCII characters and, in particular, to read UTF-8 as an encoding. Of the systems that can only render US-ASCII, only a small subset drop non-ASCII characters: the others show an incorrect character in its place. Thus, the person using such a system can often see that there is a problem, and can possibly choose to get better display software.

A.2. Difficulty in Printing

Some printers can only print a limited set of characters due to the fact that they are character-oriented, not graphical. Such printers inherently cannot print characters they do not understand. Almost all such printers print the visible ASCII characters just fine, but

many cannot print the formfeeds currently used correctly.

There are, of course, still such printers, and there always will be. However, the number is dwindling as older printers are replaced with ones that can print graphics so that now-common text features like boldface and italics can be printed.

A.3. Insufficient Fonts

Almost no display system that can display text that is encoded with UTF-8 can display every character in the Unicode repertoire. Thus, some non-ASCII characters that are included in RFCs will not display properly.

Virtually every system that can display Unicode knows how to substitute a replacement character for ones that cannot be displayed. In fact, many such systems have glyphs for rendering unknown characters and different glyphs for rendering known characters for which the system has no font.

A.4. Inability to Search for Non-ASCII Characers

If authors start using non-ASCII characters in their names and/or addresses, people who know the characters but are unfamiliar with the user interface on their computers may not be able to enter those characters in the search criteria. For example, some people do not know how to enter "u-with-umlaut" in their operating system, even though the operating system allows such input.

This is a valid concern, but one that is orthogonal to whether or not RFCs should use these characters. The alternative (never go to UTF-8) simply shifts the problem to forcing the user to guess which ASCII-only spelling to use when searching.

Appendix B. Changes from -05 to -06

None significant.

Authors' Addresses

Paul Hoffman
VPN Consortium

Email: paul.hoffman@vpnc.org

Tim Bray
Sun Microsystems

Email: tbray@textuality.com