Network Working Group
Request for Comments: 5335
Updates: 2045, 2822
Category: Experimental

Y. Abel, Ed.
TWNIC

September 2008

Internationalized Email Headers

Status of This Memo

This memo defines an Experimental Protocol for the Internet community. It does not specify an Internet standard of any kind. Discussion and suggestions for improvement are requested. Distribution of this memo is unlimited.

Abstract

Full internationalization of electronic mail requires not only the capabilities to transmit non-ASCII content, to encode selected information in specific header fields, and to use non-ASCII characters in envelope addresses. It also requires being able to express those addresses and the information based on them in mail header fields. This document specifies an experimental variant of Internet mail that permits the use of Unicode encoded in UTF-8, rather than ASCII, as the base form for Internet email header field. This form is permitted in transmission only if authorized by an SMTP extension, as specified in an associated specification. This specification Updates section 6.4 of RFC 2045 to conform with the requirements.

Table of Contents

1. Introduction

1.1. Role of This Specification

Full internationalization of electronic mail requires several capabilities:

- The capability to transmit non-ASCII content, provided for as part of the basic MIME specification [RFC2045], [RFC2046].
- The capability to use international characters in envelope addresses, discussed in [RFC4952] and specified in [RFC5336].
- The capability to express those addresses, and information related to them and based on them, in mail header fields, defined in this document.

This document specifies an experimental variant of Internet mail that permits the use of Unicode encoded in UTF-8 [RFC3629], rather than ASCII, as the base form for Internet email header fields. This form is permitted in transmission, if authorized by the SMTP extension specified in [RFC5336] or by other transport mechanisms capable of processing it.

1.2. Relation to Other Standards

This document updates Section 6.4 of RFC 2045. It removes the blanket ban on applying a content-transfer-encoding to all subtypes of message/, and instead specifies that a composite subtype MAY specify whether or not a content-transfer-encoding can be used for that subtype, with "cannot be used" as the default.

This document also updates [RFC2822] and MIME ([RFC2045]), and the fact that an Experimental specification updates a Standards-Track specification means that people who participate in the experiment have to consider those standards updated.

Allowing use of a content-transfer-encoding on subtypes of messages is not limited to transmissions that are authorized by the SMTP extension specified in [RFC5336]. Message/global permits use of a content-transfer-encoding.

2. Background and History

Mailbox names often represent the names of human users. Many of these users throughout the world have names that are not normally expressed with just the ASCII repertoire of characters, and would like to use more or less their real names in their mailbox names.

These users are also likely to use non-ASCII text in their common names and subjects of email messages, both received and sent. This protocol specifies UTF-8 as the encoding to represent email header field bodies.

The traditional format of email messages [RFC2822] allows only ASCII characters in the header fields of messages. This prevents users from having email addresses that contain non-ASCII characters. It further forces non-ASCII text in common names, comments, and in free text (such as in the Subject: field) to be encoded (as required by MIME format [RFC2047]). This specification describes a change to the email message format that is related to the SMTP message transport change described in the associated document [RFC4952] and [RFC5336], and that allows non-ASCII characters in most email header fields. These changes affect SMTP clients, SMTP servers, mail user agents (MUAs), list expanders, gateways to other media, and all other processes that parse or handle email messages.

As specified in [RFC5336], an SMTP protocol extension "UTF8SMTP" is used to prevent the transmission of messages with UTF-8 header fields to systems that cannot handle such messages.

Use of this SMTP extension helps prevent the introduction of such messages into message stores that might misinterpret, improperly display, or mangle such messages. It should be noted that using an ESMTP extension does not prevent transferring email messages with UTF-8 header fields to other systems that use the email format for messages and that may not be upgraded, such as unextended POP and IMAP servers. Changes to these protocols to handle UTF-8 header fields are addressed in [EAI-POP] and [IMAP-UTF8] .

The objective for this protocol is to allow UTF-8 in email header fields. Issues such as how to handle messages containing UTF-8 header fields that have to be delivered to systems that have not been upgraded to support this capability are discussed in [DOWNGRADE].

3. Terminology

A plain ASCII string is also a valid UTF-8 string; see [RFC3629]. In this document, ordinary ASCII characters are UTF-8 characters if they are in headers which contain <utf8-xtra-char>s.

Unless otherwise noted, all terms used here are defined in [RFC2821], [RFC2822], [RFC4952], or [RFC5336].

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 [RFC2119].

4. Changes on Message Header Fields

SMTP clients can send header fields in UTF-8 format, if the UTF8SMTP extension is advertised by the SMTP server or is permitted by other transport mechanisms.

This protocol does NOT change the [RFC2822] rules for defining header field names. The bodies of header fields are allowed to contain UTF-8 characters, but the header field names themselves must contain only ASCII characters.

To permit UTF-8 characters in field values, the header definition in [RFC2822] must be extended to support the new format. The following ABNF is defined to substitute those definitions in [RFC2822].

The syntax rules not covered in this section remain as defined in [RFC2822].

4.1. UTF-8 Syntax and Normalization

UTF-8 characters can be defined in terms of octets using the following ABNF [RFC5234], taken from [RFC3629]:

```
UTF8-xtra-char  =   UTF8-2 / UTF8-3 / UTF8-4

UTF8-2          =   %xC2-DF UTF8-tail

UTF8-3          =   %xE0 %xA0-BF UTF8-tail /
                    %xE1-EC 2(UTF8-tail) /
                    %xED %x80-9F UTF8-tail /
                    %xEE-EF 2(UTF8-tail)

UTF8-4          =   %xF0 %x90-BF 2( UTF8-tail ) /
                    %xF1-F3 3( UTF8-tail ) /
                    %xF4 %x80-8F 2( UTF8-tail )

UTF8-tail       =   %x80-BF
```

These are normatively defined in [RFC3629], but kept in this document for reasons of convenience.

See [RFC5198] for a discussion of normalization; the use of normalization form NFC is RECOMMENDED.

4.2. Changes on MIME Headers

This specification updates Section 6.4 of [RFC2045]. [RFC2045] prohibits applying a content-transfer-encoding to all subtypes of message/. This specification relaxes the rule -- it allows newly defined MIME types to permit content-transfer-encoding, and it allows content-transfer-encoding for message/global (see Section 4.6).

Background: Normally, transfer of message/global will be done in 8-bit-clean channels, and body parts will have "identity" encodings, that is, no decoding is necessary. In the case where a message containing a message/global is downgraded from 8-bit to 7-bit as described in [RFC1652], an encoding may be applied to the message; if the message travels multiple times between a 7-bit environment and an environment implementing UTF8SMTP, multiple levels of encoding may occur. This is expected to be rarely seen in practice, and the potential complexity of other ways of dealing with the issue are thought to be larger than the complexity of allowing nested encodings where necessary.

4.3. Syntax Extensions to RFC 2822

The following rules are intended to extend the corresponding rules in [RFC2822] in order to allow UTF-8 characters.

```
FWS     =  <see [RFC2822], folding white space>

CFWS    =  <see [RFC2822], folding white space>

ctext   =/  UTF8-xtra-char

utext   =/  UTF8-xtra-char

comment =   "(" *([FWS] utf8-ccontent) [FWS] ")"

word    =   utf8-atom / utf8-quoted-string
```

This means that all the [RFC2822] constructs that build upon these will permit UTF-8 characters, including comments and quoted strings. We do not change the syntax of <atext> in order to allow UTF8 characters in <addr-spec>. This would also allow UTF-8 characters in <message-id>, which is not allowed due to the limitation described in Section 4.5. Instead, <utf8-atext> is added to meet this requirement.

```
utf8-text   =  %d1-9 /          ; all UTF-8 characters except
               %d11-12 /        ; US-ASCII NUL, CR, and LF
               %d14-127 /
               UTF8-xtra-char

utf8-quoted-pair   = ("\" utf8-text) / obs-qp

utf8-qcontent      = utf8-qtext / utf8-quoted-pair

utf8-quoted-string = [CFWS]
                     DQUOTE *([FWS] utf8-qcontent) [FWS] DQUOTE
                     [CFWS]

utf8-ccontent =     ctext / utf8-quoted-pair / comment

utf8-qtext    =     qtext / UTF8-xtra-char

utf8-atext   =  ALPHA / DIGIT /
                "!" / "#" /     ; Any character except
                "$" / "%" /     ; controls, SP, and specials.
                "&" / "'" /     ; Used for atoms.
                "*" / "+" /
                "-" / "/" /
                "=" / "?" /
                "^" / "_" /
                "`" / "{" /
                "|" / "}" /
                "~" /
                UTF8-xtra-char

utf8-atom     = [CFWS] 1*utf8-atext [CFWS]

utf8-dot-atom = [CFWS] utf8-dot-atom-text [CFWS]

utf8-dot-atom-text = 1*utf8-atext *("." 1*utf8-atext)

qcontent      = utf8-qcontent
```

To allow the use of UTF-8 in a Content-Description header field [RFC2045], the following syntax is used:

```
description   = "Content-Description:" unstructured CRLF
```

The <utext> syntax is extended above to allow UTF-8 in all <unstructured> header fields.

Note, however, this does not remove any constraint on the character set of protocol elements; for instance, all the allowed values for timezone in the Date: headers are still expressed in ASCII. And also, none of this revised syntax changes what is allowed in a <msg-id>, which will still remain in pure ASCII.

4.4. Change on addr-spec Syntax

Internationalized email addresses are represented in UTF-8. Thus, all header fields containing <mailbox>es are updated to permit UTF-8 as well as an additional, optional all-ASCII alternate address. Note that Message Submission Servers ("MSAs") and Message Transfer Agents (MTAs) may downgrade internationalized messages as needed. The procedure for doing so is described in [DOWNGRADE].

```
mailbox        =  name-addr / addr-spec / utf8-addr-spec

angle-addr     =/ [CFWS] "<" utf8-addr-spec [ alt-address ] ">"
                  [CFWS] / obs-angle-addr

utf8-addr-spec =  utf8-local-part "@" utf8-domain

utf8-local-part=  utf8-dot-atom / utf8-quoted-string / obs-local-part

utf8-domain    =  utf8-dot-atom / domain-literal / obs-domain

alt-address    =  FWS "<" addr-spec ">"
```

Below are a few examples of possible <mailbox> representations.

```
"DISPLAY_NAME" <ASCII@ASCII>
   ; traditional mailbox format

"DISPLAY_NAME" <non-ASCII@non-ASCII>
   ; UTF8SMTP but no ALT-ADDRESS parameter provided,
   ; message will bounce if UTF8SMTP extension is not supported

<non-ASCII@non-ASCII>
   ; without DISPLAY_NAME and quoted string
   ; UTF8SMTP but no ALT-ADDRESS parameter provided,
   ; message will bounce if UTF8SMTP extension is not supported

"DISPLAY_NAME" <non-ASCII@non-ASCII <ASCII@ASCII>>
   ; UTF8SMTP with ALT-ADDRESS parameter provided,
   ; ALT-ADDRESS can be used if downgrade is necessary
```

4.5. Trace Field Syntax

"For" fields containing internationalized addresses are allowed, by use of the new uFor syntax. UTF-8 information may be needed in Received fields. Such information is therefore allowed to preserve the integrity of those fields. The uFor syntax retains the original UTF-8 email address between email address internationalization (EAI)-aware MTAs. Note that, should downgrading be required, the uFor parameter is dropped per the procedure specified in [DOWNGRADE].

The "Return-Path" header provides the email return address in the mail delivery. Thus, the header is augmented to carry UTF-8 addresses (see the revised syntax of <angle-addr> in Section 4.4 of this document). This will not break the rule of trace field integrity, because the header is added at the last MTA and described in [RFC2821].

The <item-value> on "Received:" syntax is augmented to allow UTF-8 email address in the "For" field. <angle-addr> is augmented to include UTF-8 email address. In order to allow UTF-8 email addresses in an <addr-spec>, <utf8-addr-spec> is added to <item-value>.

```
item-value     =/      utf8-addr-spec
```

4.6. message/global

Internationalized messages must only be transmitted as authorized by [RFC5336] or within a non-SMTP environment which supports these messages. A message is a "message/global message", if

- it contains UTF-8 header values as specified in this document, or

- it contains UTF-8 values in the headers fields of body parts.

The type message/global is similar to message/rfc822, except that it contains a message that can contain UTF-8 characters in the headers of the message or body parts. If this type is sent to a 7-bit-only system, it has to be encoded in MIME [RFC2045]. (Note that a system compliant with MIME that doesn’t recognize message/global would treat it as "application/octet-stream" as described in Section 5.2.4 of [RFC2046].)

Alternatively, SMTP servers and other systems which transfer a message/global body part MAY choose to down-convert it to a message/rfc822 body part using the rules described in [DOWNGRADE].

Type name: message

Subtype name: global

Required parameters: none

Optional parameters: none

Encoding considerations: Any content-transfer-encoding is permitted. The 8-bit or binary content-transfer-encodings are recommended where permitted.

Security considerations: See Section 5.

Interoperability considerations: The media type provides functionality similar to the message/rfc822 content type for email messages with international email headers. When there is a need to embed or return such content in another message, there is generally an option to use this media type and leave the content unchanged or down-convert the content to message/rfc822. Both of these choices will interoperate with the installed base, but with different properties. Systems unaware of international headers will typically treat a message/global body part as an unknown attachment, while they will understand the structure of a message/rfc822. However, systems that understand message/global will provide functionality superior to the result of a down-conversion to message/rfc822. The most interoperable choice depends on the deployed software.

Published specification: RFC 5335

Applications that use this media type: SMTP servers and email clients that support multipart/report generation or parsing. Email clients which forward messages with international headers as attachments.

Additional information:

Magic number(s): none

File extension(s): The extension ".u8msg" is suggested.

Macintosh file type code(s): A uniform type identifier (UTI) of "public.utf8-email-message" is suggested. This conforms to "public.message" and "public.composite-content", but does not necessarily conform to "public.utf8-plain-text".

Person & email address to contact for further information: See the Author's Address section of this document.

Intended usage: COMMON

Restrictions on usage: This is a structured media type which embeds other MIME media types. The 8-bit or binary content-transfer-encoding MUST be used unless this media type is sent over a 7-bit-only transport.

Author: See the Author's Address section of this document.

Change controller: IETF Standards Process

5. Security Considerations

If a user has a non-ASCII mailbox address and an ASCII mailbox address, a digital certificate that identifies that user may have both addresses in the identity. Having multiple email addresses as identities in a single certificate is already supported in PKIX (Public Key Infrastructure for X.509 Certificates) and OpenPGP.

Because UTF-8 often requires several octets to encode a single character, internationalized local parts may cause mail addresses to become longer. As specified in [RFC2822], each line of characters MUST be no more 998 octets, excluding the CRLF.

Because internationalized local parts may cause email addresses to be longer, processes that parse, store, or handle email addresses or local parts must take extra care not to overflow buffers, truncate addresses, or exceed storage allotments. Also, they must take care, when comparing, to use the entire lengths of the addresses.

In this specification, a user could provide an ASCII alternative address for a non-ASCII address. However, it is possible these two addresses go to different mailboxes, or even different people. This configuration may be based on a user's personal choice or on administration policy. We recognize that if ASCII and non-ASCII email is delivered to two different destinations, based on MTA capability, this may violate the principle of least astonishment, but this is not a "protocol problem".

The security impact of UTF-8 headers on email signature systems such as Domain Keys Identified Mail (DKIM), S/MIME, and OpenPGP is discussed in RFC 4952, Section 9. A subsequent document [DOWNGRADE] will cover the impact of downgrading on these systems.

6. IANA Considerations

IANA has registered the message/global MIME type using the registration form contained in Section 4.4.

7. Acknowledgements

This document incorporates many ideas first described in Internet-Draft form by Paul Hoffman, although many details have changed from that earlier work.

The author especially thanks Jeff Yeh for his efforts and contributions on editing previous versions.

Most of the content of this document is provided by John C Klensin. Also, some significant comments and suggestions were received from Charles H. Lindsey, Kari Hurtta, Pete Resnick, Alexey Melnikov, Chris Newman, Yangwoo Ko, Yoshiro Yoneya, and other members of the JET team (Joint Engineering Team) and were incorporated into the document. The editor sincerely thanks them for their contributions.

8. References

8.1. Normative References

[RFC1652] Klensin, J., Freed, N., Rose, M., Stefferud, E., and D. Crocker, "SMTP Service Extension for 8bit-MIMEtransport", RFC 1652, July 1994.

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

[RFC2821] Klensin, J., "Simple Mail Transfer Protocol", RFC 2821, April 2001.

[RFC2822] Resnick, P., "Internet Message Format", RFC 2822, April 2001.

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

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

[RFC5198] Klensin, J. and M. Padlipsky, "Unicode Format for Network Interchange", RFC 5198, March 2008.

[RFC5234] Crocker, D. and P. Overell, "Augmented BNF for Syntax Specifications: ABNF", STD 68, RFC 5234, January 2008.

[RFC5336] Yao, J., Ed. and W. Mao, Ed., "SMTP Extension for Internationalized Email Addresses", RFC 5336, September 2008.

8.2. Informative References

[DOWNGRADE] Fujiwara, K. and Y. Yoneya, "Downgrading mechanism for Email Address Internationalization", Work in Progress, July 2008.

[EAI-POP] Newman, C. and R. Gellens, "POP3 Support for UTF-8", Work in Progress, July 2008.

[IMAP-UTF8] Resnick, P. and C. Newman, "IMAP Support for UTF-8", Work in Progress, April 2008.

[RFC2045] Freed, N. and N. Borenstein, "Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies", RFC 2045, November 1996.

[RFC2046] Freed, N. and N. Borenstein, "Multipurpose Internet Mail Extensions (MIME) Part Two: Media Types", RFC 2046, November 1996.

[RFC2047] Moore, K., "MIME (Multipurpose Internet Mail Extensions) Part Three: Message Header Extensions for Non-ASCII Text", RFC 2047, November 1996.

Author’s Address

Abel Yang (editor)
TWNIC
4F-2, No. 9, Sec 2, Roosvelt Rd.
Taipei, 100
Taiwan

Phone: +886 2 23411313 ext 505
EMail: abelyang@twnic.net.tw

Full Copyright Statement

Intellectual Property