

**Internationalized Email Headers
draft-ietf-eai-utf8headers-03.txt**

Status of this Memo

By submitting this Internet-Draft, each author represents that any applicable patent or other IPR claims of which he or she is aware have been or will be disclosed, and any of which he or she becomes aware will be disclosed, in accordance with [Section 6 of 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 August 14, 2007.

Copyright Notice

Copyright (C) The IETF Trust (2007).

Abstract

Full internationalization of electronic mail requires not only the capability 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 information based on them in mail header fields. This document specifies the use of Unicode encoded in UTF-8, rather than ASCII, as the base form for Internet email header field bodies. This form is permitted in transmission only if authorized by an SMTP extension, as specified in an associated specification.

Table of Contents

1.	Introduction	3
1.1.	Role of this specification	3
2.	Background and History	3
3.	Terminology	4
4.	Pre-requirement	4
5.	Changes on Message Header Fields	5
5.1.	UTF8 Syntax	5
5.2.	Syntax extensions to RFC 2822	6
5.3.	Change on addr-spec syntax	8
5.4.	Trace field syntax	8
6.	Additional issues	9
6.1.	Mailing list header fields	9
6.2.	MIME headers	9
7.	Security Considerations	9
8.	IANA considerations	10
9.	Acknowledgements	10
10.	Edit history	10
10.1.	draft-ietf-eai-utf8header-02 => draft-ietf-eai-utf8header-03	10
10.2.	draft-ietf-eai-utf8header-01 => draft-ietf-eai-utf8header-02	11
10.3.	draft-ietf-eai-utf8header-00 => draft-ietf-eai-utf8header-01	11
10.4.	draft-yeh-ima-utf8header-01 => draft-ietf-eai-utf8header-00	11
10.5.	draft-yeh-ima-utf8header-00 => draft-yeh-ima-utf8header-01	11
11.	References	11
11.1.	Normative References	11
11.2.	Informative References	12
	Author's Address	13
	Intellectual Property and Copyright Statements	14

Yeh

Expires August 14, 2007

[Page 2]

1. Introduction

1.1. Role of this specification

Full internationalization of electronic mail requires several capabilities:

- o The capability to transmit non-ASCII content, provided for as part of the basic MIME specification [[RFC2045](#)], [[RFC2046](#)].
- o The capability to encode selected information in specific header fields, provided for as another part of the MIME specification [[RFC2047](#)].
- o The capability to use international characters in envelope addresses, discussed in [[EAI-overview](#)] and specified in [[EAI-SMTP-extension](#)]. And, finally,
- o The capability to express those addresses, and information related to and based on them, in mail header fields, defined in this document.
- o The capability to use international characters in other headers, but only as expressly permitted herein, or in future extensions.

This document specifies 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 [[EAI-SMTP-extension](#)] or by other transport mechanisms capable of processing it.

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 represented with just the ASCII repertoire of characters, and would more or less like to use 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 in what they send and what they receive. 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 in 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 specifications [[EAI-overview](#)] and [[EAI-SMTP-extension](#)], and that allows non-ASCII characters throughout email header fields.

Yeh

Expires August 14, 2007

[Page 3]

These changes affect SMTP clients, SMTP servers, mail user agents (MUAs), list expanders and gateways to other media.

As specified in [[EAI-SMTP-extension](#)], 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 against the introduction of such messages into message stores that might misrepresent or mangle such messages. It should be noted that using an ESMTP extension does not prevent against 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 the POP and IMAP protocols. Those protocols also need to be changed in order to handle stored messages that have UTF-8 header fields.

The objective for this protocol is to allow UTF-8 in email header fields. Issues about how to handle messages that contain UTF-8 header fields but are proposed to be delivered to systems that have not been upgraded to support this capability are discussed elsewhere, particularly in [[EAI-downgrading](#)].

3. Terminology

In this document, header fields are "UTF-8 headers" if the bodies of those headers contain UTF-8 characters.

Unless otherwise noted, all terms used here are defined in [[RFC2821](#)] or [[RFC2822](#)] or in [[EAI-overview](#)].

The key words "MUST", "SHALL", "REQUIRED", "SHOULD", "RECOMMENDED", and "MAY" in this document are to be interpreted as described in [RFC 2119](#) [[RFC2119](#)].

This document is being discussed on the ima mailing list. See <https://www1.ietf.org/mailman/listinfo/ima> for information about subscribing. The list's archive is at <http://www1.ietf.org/mail-archive/web/ima/index.html>.

4. Pre-requirement

The use of UTF-8 header fields is dependent on the use of an SMTP extension named "UTF8SMTP" or of similar capabilities in other transports.

That protocol is defined in [[EAI-SMTP-extension](#)]. If that extension

Yeh

Expires August 14, 2007

[Page 4]

is not supported, UTF-8 header fields MUST NOT be transmitted by SMTP.

Sending MUAs conforming to this specification MUST encode all header fields in UTF-8. No other direct encodings (like Big-5) are allowed. Although there is nothing wrong with the continued use of [\[RFC2047\]](#), it is not recommended in this document.

5. Changes on Message Header Fields

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

This protocol does NOT change the definition of header field names. That is, only the bodies of header fields are allowed to have UTF-8 characters; the rules in [RFC 2822](#) for header names are not changed. To be able to do so, the header definition in [RFC 2822](#) must be extended to support new format. The following ABNF is defined to substitute those definitions in [RFC 2822](#).

For those syntax rules not referred to in this section, the original definition in [RFC 2822](#).

5.1. UTF8 Syntax

The use of UTF8 characters are defined as following.

```
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-F7 3(UTF8-tail)

UTF8-tail       = %x80-BF
```

These are taken from RFC 3629, but kept in this document for reasons of convenience.

[Note in draft: Whether normalizing is needed or not will be placed here.]

Yeh

Expires August 14, 2007

[Page 5]

5.2. Syntax extensions to [RFC 2822](#)

The following rules are intended to extend the corresponding rules in [RFC 2822](#) to allow UTF8 characters.

```
ctext    = NO-WS-CTL /      ; all of <text> except
              %d33-39 /      ; SP, HTAB, "(", ")"
              %d42-91 /      ; and "\"
              %d93-126 /
              UTF8-xtra-char

utext     = NO-WS-CTL /      ; Non white space controls
              %d33-126 /      ; The rest of US-ASCII
              UTF8-xtra-char

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

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

This means that all the [RFC 2822](#) constructs that build upon these will permit UTF-8 characters, including comments and quoted strings. Besides, in order to allow UTF8 characters in <addr-spec> we have to change the syntax of <atext>. However, it will also lead <msg-id> to allow UTF8 characters, which is not allowed due to the limitation described in [Section 5.4](#). So <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    = NO-WS-CTL /      ; all of <text> except
                %d33 /           ; The rest of the US-ASCII
                %d35-91 /        ; characters not including "\"
                %d93-126 /       ; or the quote character
                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)
```

[NOTE IN DRAFT: If any header needs to be restricted to disallow this, please raise the issue on the mailing list.]

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 affects what is allowed in a <message-id>, which will still remain in pure ASCII.

Yeh

Expires August 14, 2007

[Page 7]

5.3. Change on addr-spec syntax

In this specification, internationalized email address will be presented in UTF-8. Thus, all header fields involving <mailbox>es may be different from traditional ones. There might be UTF8SMTP unaware MTAs in the mail routing path. In that case, MTA may bounce the message with reply code 550, or downgrade the non-ASCII contents of all header bodies before continuing to send the message. The downgrade process involve with a new ALT-ADDRESS parameter. When downgrade occurs, the ALT-ADDRESS will be used for mail delivery instead of the internationalized email address, the detail is described in [[EAI-downgrading](#)].

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

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

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       = [CFWS] "<" addr-spec ">" [CFWS]
```

Below list a few possible <mailbox> representation as example.

```
"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

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

5.4. Trace field syntax

Internationalized domain names in Received fields must be transmitted in punycode form. "For" fields containing internationalized addresses are prohibited, since subsequent downgrading would force violating rules in [RFC 2821](#) prohibiting altering existing Received fields. With these two restrictions, there should be no need for UTF-8 information in Received fields and such information is prohibited to preserve the integrity of those fields. More

Yeh

Expires August 14, 2007

[Page 8]

generally, UTF-8 information of any sort MUST NOT appear in Received fields, even in comments within those fields.

The "Return-Path" header provides the email returning address in the mail delivery. Thus, it MUST be able to carry UTF8 addresses (see the revised syntax of <angle-addr> in [Section 5.2](#) of this document). This will not break the rule of trace field integrity, because it is added at the last MTA.

6. Additional issues

This section identifies issues that are not covered as part of this set of specifications, but that will need to be considered as part of UTF8SMTP deployment.

6.1. Mailing list header fields

All mailing list and mail redistribution related header are discussed in [[EAI-mailing-list](#)].

6.2. MIME headers

The syntax of <value>, as defined in [RFC 2045](#), is

value = token / quoted-string

To be able to use UTF-8 characters in MIME headers, <quoted-string> syntax is extended as

qcontent = utf8-qtext / quoted-pair

In all those headers, such as Content-Type and Content-Disposition [plus lots of others being defined in various other documents], which make use of <value> within <parameter> as defined in [[RFC2045](#)] as modified by [[RFC2231](#)], it will now be allowed to use <quoted-string>s containing UTF-8 characters (see the revised syntax of <utf8-qtext> in [Section 5.2](#) of this document).

7. 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 and OpenPGP.

Yeh

Expires August 14, 2007

[Page 9]

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 [RFC 2822](#), each line of characters MUST be no more 998 octets, excluding the CRLF.

In this specification, a user could provide an ASCII alternative address for a non-ASCII address. However, it is possible these two address go to different mailbox, or even different persons. This might not be a protocol problem, but the user's personal choice or administration policy or even be a deliberate attempt to deceive or cause confusion.

8. IANA considerations

There is no IANA considerations in this document.

9. Acknowledgements

This document was created by incorporating a good deal of material from an old Internet Draft by Paul Hoffman [[Hoffman-utf8-headers](#)]. While many of the concepts and details have changed, the contributions from that draft are greatly appreciated.

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, Chris Newman, Yangwoo KO, Yoshiro YONEYA, and other members of the JET team and were incorporated into the document. The editor is much great thanks to their contribution sincerely.

10. Edit history

This section is used for tracking the update of this document. Will be removed after finalize.

10.1. [draft-ietf-eai-utf8header-02](#) => [draft-ietf-eai-utf8header-03](#)

1. Editrial changes on terms and english.
2. ABNF revise.
3. addr-spec change, put ALT-ADDRESS inside "<" and ">" quote with "<" and ">".
4. Remove the "Header-Type" header.
5. Remove the content in IANA considerations since "Header-Type" is removed.

Yeh

Expires August 14, 2007

[Page 10]

10.2. [draft-ietf-eai-utf8header-01](#) => [draft-ietf-eai-utf8header-02](#)

1. Editorial changes on terms and english.
2. Change the header name "UTF8SMTP" to "Header-Type", and ABNF revise.
3. addr-spec change, put ALT-ADDRESS inside "<" and ">" quote with "[" and "]".
4. IANA considerations section rewrite into registration templates specified in [RFC 3864](#).

10.3. [draft-ietf-eai-utf8header-00](#) => [draft-ietf-eai-utf8header-01](#)

1. ABNF revise.
2. Terminology sync with overview document.
3. addr-spec change, put ALT-ADDRESS inside "<" and ">" quote with "{" and "}".
4. add IANA considerations to register the new 2822 header "UTF8SMTP".
5. add Security considerations about relation of UTF8SMTP address to ALT-ADDRESS.

10.4. [draft-yeh-ima-utf8header-01](#) => [draft-ietf-eai-utf8header-00](#)

1. ABNF added.
2. Editorial changes.
3. Sent it as WG document.

10.5. [draft-yeh-ima-utf8header-00](#) => [draft-yeh-ima-utf8header-01](#)

1. Section re-arranged.
2. Remove content are not below to this document.

11. References**11.1.** Normative 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.

[EAI-SMTP-extension]

Yao, J., Ed. and Wei. Mao, "SMTP extension for internationalized email address",

Yeh

Expires August 14, 2007

[Page 11]

[draft-ietf-eai-smtpext-02.txt](#) (work in progress),
July 2006.

[EAI-mailing-list]

Gellens, Randall., "Mailing Lists and Internationalized Email Addresses", [draft-ietf-eai-mailinglist-01.txt](#) (work in progress), January 2007.

[EAI-overview]

Klensin, J. and Y. Ko, "Overview and Framework of Internationalized Email Address Delivery",
[draft-ietf-eai-framework-05.txt](#) (work in progress),
February 2007.

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

[RFC2231] Freed, N. and K. Moore, "MIME Parameter Value and Encoded Word Extensions: Character Sets, Languages, and Continuations", [RFC 2231](#), November 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.

[11.2. Informative References](#)

[EAI-downgrading]

YONEYA, Yoshiro., Ed. and Kazunori. Fujiwara, Ed.,
"Downgrading mechanism for Internationalized eMail Address (IMA)", [draft-ietf-eai-downgrade-01.txt](#) (work in progress), June 2006.

[Hoffman-utf8-headers]

Hoffman, P., "SMTP Service Extensions or Transmission of Headers in UTF-8 Encoding",
[draft-hoffman-utf8headers-00.txt](#) (work in progress),
December 2003.

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

Yeh

Expires August 14, 2007

[Page 12]

- [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.
- [RFC3864] Klyne, G., Nottingham, M., and J. Mogul, "Registration Procedures for Message Header Fields", [BCP 90](#), [RFC 3864](#), September 2004.
- [RFC4646] Phillips, A. and M. Davis, "Tags for Identifying Languages", [BCP 47](#), [RFC 4646](#), September 2006.

Author's Address

Jeff Yeh (editor)
TWNIC
4F-2, No. 9, Sec 2, Roosevelt Rd.
Taipei, 100
Taiwan

Phone: +886 2 23411313 ext 506
Email: jeff@twNIC.net.tw

Full Copyright Statement

Copyright (C) The IETF Trust (2007).

This document is subject to the rights, licenses and restrictions contained in [BCP 78](#), and except as set forth therein, the authors retain all their rights.

This document and the information contained herein are provided on an "AS IS" basis and THE CONTRIBUTOR, THE ORGANIZATION HE/SHE REPRESENTS OR IS SPONSORED BY (IF ANY), THE INTERNET SOCIETY, THE IETF TRUST AND THE INTERNET ENGINEERING TASK FORCE DISCLAIM ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE INFORMATION HEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Intellectual Property

The IETF takes no position regarding the validity or scope of any Intellectual Property Rights or other rights that might be claimed to pertain to the implementation or use of the technology described in this document or the extent to which any license under such rights might or might not be available; nor does it represent that it has made any independent effort to identify any such rights. Information on the procedures with respect to rights in RFC documents can be found in [BCP 78](#) and [BCP 79](#).

Copies of IPR disclosures made to the IETF Secretariat and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementers or users of this specification can be obtained from the IETF on-line IPR repository at <http://www.ietf.org/ipr>.

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights that may cover technology that may be required to implement this standard. Please address the information to the IETF at ietf-ipr@ietf.org.

Acknowledgment

Funding for the RFC Editor function is provided by the IETF Administrative Support Activity (IASA).

Yeh

Expires August 14, 2007

[Page 14]