              Internationalized Email Headers
              draft-ietf-eai-utf8headers-00.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 December 1, 2006.

Copyright Notice

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**

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.

   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].

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 the 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 body.

   The traditional format of email messages [RFC2822] only allows 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 connected 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. These changes affect SMTP clients, SMTP servers, and
   mail user agents (MUAs).

As specified in [EAI-SMTP-extension], an SMTP protocol extension
[RFC2821] is used to prevent the transmission of messages with UTF-8
header fields to systems that cannot handle such messages.

Use 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 will 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].

This protocol is workable even if internationalized email addresses
are not presented. For example, the protocol might still be used if
just the subject header has non-ASCII characters, but the protocol
MUST be used if other header fields (particularly trace header fields
such as "Received:") contain non-ASCII characters.

3. Terminology

In this document, header fields are "UTF-8 header" if the bodies of
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 "i-Email".

That protocol is defined in [EAI-SMTP-extension]. If that extension
is not supported, UTF-8 header fields MUST NOT be transmitted.

Sending MUAs that follow this protocol MUST create all header fields
encoded in UTF-8. No other direct encodings are allowed.

5. Identification of internationalized email

When a SMTP client tries to send a mail to a SMTP server that does
not support i-Email, the client should know whether the message
requires the support for i-Email or not. In addition to this,
identification of internationalized email is also required when a
message is stored and resent. Checking the presence of UTF-8
characters in the header whenever such an identification is required
may also achieve the goal. However, this type of repeated processing
wastes time and processing power of involved systems. It is nice to
have a mechanism (such as self-label) or some indicator to identify
whether the message is new format(i.e. i-Email compliant) or old one
(i.e. RFC 2822 compliant).

To be able to do so, sending MUA MUST insert a new header field to
identify the presence of i18n information (particularly UTF-8
headers) in the message. The new header specified as "i-Email", and
elements of the header is the version number of i18n email. The i18n
header field syntax specified like:

i-Email: 1.0

[Note in draft: There should be more useful information can be place
in the new header field.]

While we can't require ordering of headers, it would be good to have
it appear as near the top of the headers as possible. It would also
be good to be able to guarantee that it will be there when the
message is dropped into a mail store. Thus, when a i18n email is
delivered.

o The "i-Email" header field MUST be inserted by the originating
   MUA.
o The "i-Email" header field MUST be inserted, along with Return-
   path, by the final delivery MTA if not presented.
o The "i-Email" header field, if present, MUST be removed as part of
   any downgrading process that eliminates the UTF-8 header
   information.
o MTAs MAY check for duplicates of the "i-Email" header field and
   eliminate all but one of them. However, if a receiving MUA
   encounters more than one of these headers, it SHOULD simply ignore

any excess ones.

This combination guarantees that the header will be present on delivery even if it is deleted in transit.

6. **Changes on Message Header Fields**

SMTP client can send header fields in UTF-8 format, if the IEmail extension advertised by SMTP server. However, the Message-ID is the unique identifier of a single email. [Note in draft: Extension name depends on the SMTP extension defined in [EAI-SMTP-extension]]

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 extended to support new format. That following ABNF is defined to substitute those definition in RFC 2822.

For those tokens not referred in this section remains as the original definition in RFC 2822.

6.1. **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 FRC 3629, but keep in this document for convenient reason.
[Note in draft: Weather normalizing is needed or not will be place in here.]

6.2. Syntax extend from RFC 2822

   The following rules are intended to supersede the corresponding rules
   in RFC 2822.

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

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

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

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

unstructured = 1*( [FWS] utext ) [FWS]

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

6.3. 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 i-Email
   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, as
described in [EAI-downgrading]. However, MTAs never know if there
are any instructions or data embedded in the email address (such as
'+','!','%',...) or not. The only one way is to let the mail address
owner to tell if the address is ok for downgrade process. Hence, the
ATOMIC and ALT-ADDRESS options are introduced. The detail of ATOMIC
and ALT-ADDRESS options can be found in [EAI-SMTP-extension].

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

alt-address    =  alt-separator stct-addr /
                  alt-separator "atomic"

alt-separator  =  [FWS] "," [FWS]
```

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

```
   DISPLAY NAME <ASCII@DOMAIN>
      ; tradition mailbox format

   DISPLAY NAME <IEMAIL@IDNA>
      ; i-Email but no ALT-ADDRESS nor ATOMIC option provided,
      ; message will bounce if i-Email extension is not supported

   DISPLAY NAME <IEMAIL@IDNA , ATOMIC>
      ; i-Email with ATOMIC option provided
      ; message is good for downgrade

   DISPLAY NAME <IEMAIL@IDNA , ASCII@DOMAIN>
      ; i-Email with ALT-ADDRESS provided
      ; ALT-ADDRESS can be used if downgrade is necessary
```

6.4. ASCII address syntax

<stct-addr> will be more or less the <addr-spec> as currently defined
in RFC 2822. The syntax follows - it is extremely tedious, because
it has to undo all the changes made to intorduce UTF-8, but here it
is, anyway.

```
   stct-addr           =    stct-local-part "@" stct-domain
   stct-local-part     =    stct-dot-atom / stct-quoted-string
   stct-dot-atom       =    [CFWS] stct-dot-atom-text [CFWS]
   stct-dot-atom-text  =    1*stct-atext *( "." 1*stct-atext 0
   stct-atext          =    ALPHA / DIGIT /
                            "!" / "#" /      ; Any character except
                            "$" / "%" /      ; controls, SP, and
                            "&" / "'" /      ; specials. Used for
                            "*" / "+" /      ; atoms
                            "-" / "/" /
                            "=" / "?" /
                            "^" / "_" /
                            "`" / "{" /
                            "|" / "}" /
                            "~"

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

   stct-qcontent       =    stct-qtext / stct-quoted-pair

   stct-qtext          =    NO-WS-CTL /      ; qtext restricted to
                            %d33 /           ; US-ASCII
                            %d35-91 /
                            %d93-126
   stct-quoted-pair    =    "\" stct-text

   stct-text           =    %d1-9 /          ; text restricted to
                            %d11-12 /        ; US-ASCII
                            %d14-127
   stct-domain         =    stct-dot-atom / domain-literal
```

6.5. Trace field syntax

   There had been discusses about the trace field (such as For,
   Received, Return-path...) on the mailing list. And also, the trace
   field is very much related to the downgrade process. This section
   will be filled after concense being made.

7. Additional issue

   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
   i-Email deployment.

7.1. Mailing list header fields

   All mailing list and mail redistribution related header fields may
   need further investigation.

8. Security Considerations

   If a user has a non-ASCII mailbox address and a all-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.

   Because UTF-8 often requires several octets to encode a single
   character, internationalized local parts may cause mail addresses to
   become longer. Then may possibly make it harder to keep lines in a
   header under 78 octets. Lines that are longer than 78 octets (which
   is a SHOULD specification, not a MUST specification, in RFC 2822)
   could possibly cause mail user agents to fail in ways that affect
   security.

9. IANA considerations

   The ESMTP extension needed to support this specification is specified
   in [EAI-SMTP-extension]. This specification does not require any
   additional IANA actions in that regard.

10. 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, 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.

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",
                draft-ietf-eai-smtpext-00.txt (work in progress),
                May 2006.

   [EAI-overview]
                Klensin, J. and Y. Ko, "Overview and Framework of
                Internationalized Email Address Delivery",
                draft-ietf-eai-framework-00.txt (work in progress),
                May 2006.

   [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.

   [RFC3066]   Alvestrand, H., "Tags for the Identification of
                Languages", BCP 47, RFC 3066, January 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-yoneya-ima-downgrade-01.txt (work in
                progress), March 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.

[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

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

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

Intellectual Property Statement

Disclaimer of Validity

Copyright Statement

Acknowledgment