SMTP extension for internationalized email address
draft-ietf-eai-smtpext-01.txt

Status of this Memo

Copyright Notice

Abstract

Internationalized email address includes two parts, the local part
and the domain part. The way email addresses are used by protocols
are different from the way domain names are used. The most critical
difference is that emails are delivered through a chain of peering
clients and servers while domain names are resolved by name servers
by looking up their own tables. In addition to this, email transport
protocols SMTP and ESMTP provide a negotiation mechanism through
which clients can make decisions for further processing. So

   internationalized email address is different from the
   internationalized domain name (IDN). It can be solved by exploiting
   the negotiation mechanism while IDN can not use the negotiation
   mechanism. So internationalized email address SHOULD be solved in
   the mail transport-level using the negotiation mechanism, which is an
   architecturally desirable approach. This document specifies the use
   of SMTP extension for internationalized email address delivery. It
   also mentions the backward compatible mechanism for downgrade
   procedure, as specified in an associated specification. The protocol
   proposed here is MTA-level solution which is feasible,
   architecturally more elegant, and not as difficult to deploy in
   relevant communities.

Table of Contents

1. Introduction

1.1. Role of this specification

   An overview document [EAI-overview] specifies the requirements for,
   and components of, full internationalization of electronic mail.
   This document specifies an element of that work, specifically the
   definition of an SMTP extension [RFC1869] for the internationalized
   email address transport delivery.

1.2. Proposal Context

   In order to use internationalized email addresses, we need to
   internationalize both the domain part and the local part of the email
   address. Domain part of the email address has been internationalized
   through IDNA [RFC3490]. But the local part of the email address
   still remains as non-internationalized.

   The syntax of Internet email addresses is restricted to a subset of
   7-bit ASCII for the domain-part, with a less-restricted subset for
   the local-part. These restrictions are specified in RFC 2821
   [RFC2821]. To be able to deliver internationalized email through
   SMTP servers, we need to upgrade SMTP server to be able to carry the
   internationalized email address. Since older SMTP servers and the
   mail-reading clients and other systems that are downstream from them
   MAY not be prepared to handle these extended addresses, an SMTP
   extension is specified to identify and protect the addressing
   mechanism.

   This specification describes a change to the email transport
   mechanism that permits non-ASCII address in both the envelope and
   header fields of messages. The context for the change is described
   in [EAI-overview] and the details of the header changes are described
   in [EAI-utf8header].

1.3. Terminology

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

   All specialized terms used in this specification are defined in the
   EAI overview [EAI-overview] or in [RFC2821] and [RFC2822]. The terms
   "ASCII address", "internationalized email address", "non-ASCII
   address", "i18mail address", "UTF8SMTP", "message" and "mailing list"
   are used with the definitions from the EAI overview document.

   This document is being discussed on the EAI 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.

2. **Mail Transport-level Protocol**

2.1. **Framework for the Internationalization Extension**

The following service extension is defined:

1. The name of the SMTP service extension is "Internationalized
    Email and Extensions";
2. The EHLO keyword value associated with this extension is
    "UTF8SMTP";
3. No parameter values are defined for this EHLO keyword value. In
    order to permit future (although unanticipated) extensions, the
    EHLO response MUST NOT contain any parameters for that keyword.
    If a parameter appears, the SMTP client that is conformant to
    this version of this specification MUST treat the ESMTP response
    as if the "UTF8SMTP" keyword did not appear.
4. An optional parameters is added to the SMTP MAIL and RCPT
    commands. The parameter is named as ALT-ADDRESS. The "ALT-
    ADDRESS" requires an all-ASCII address as a substitute for the
    i18mail addresses that we call the primary address; you can learn
    more in [EAI-overview] or [EAI-downgrading]. The value of "ALT-
    ADDRESS" is set by the sender when MUA and the submit ion server
    have a communication.
5. No additional SMTP verbs are defined by this extension.
6. Servers offering this extension MUST provide support for, and
    announce, the 8BITMIME extension [RFC1652].

2.2. **The Address Internationalization Service Extension**

An SMTP Server that announces this extension MUST be prepared to
accept a UTF-8 string [RFC3629] in any position in which RFC 2821
specifies that a "mailbox" MAY appear. That string MUST be parsed
only as specified in RFC 2821, i.e., by separating the mailbox into
source route, local part and domain part, using only the characters
colon (U+003A), comma (U+002C), and at-sign (U+0040) as specified
there. Once isolated by this parsing process, the local part MUST be
treated as opaque unless the SMTP Server is the final delivery MTA.
Any domain names that are to be looked up in the DNS MUST first be
processed into the form as specified in IDNA [RFC3490] by means of
the ToASCII() operation unless they are already in that form. Any
domain names that are to be compared to local strings SHOULD be
checked for validity and then MUST be compared as specified in
section 3.4 of IDNA.

An SMTP Client that receives the UTF8SMTP extension keyword in
response to the "EHLO" command MAY transmit a mailbox name as an
internationalized string in UTF-8 form and MAY send an
internationalized mail header [EAI-utf8header]. It MAY transmit the
domain part of that string in either punycode (derived from the IDNA
process) or UTF-8 form. If it sends the domain in UTF-8 form, the
original SMTP client SHOULD first verify that the string is valid for
a domain name according to IDNA rules. As required by RFC 2821, it
MUST not attempt to parse, evaluate, or transform the local part in
any way if the UTF8SMTP SMTP extension is offered by the server. If
the UTF8SMTP SMTP extension is not offered by the Server, the SMTP
Client MUST NOT transmit an internationalized address and MUST NOT
transmit a mail body which contains internationalized mail headers
[EAI-utf8header]. Instead, it MUST either return the message to the
user as undeliverable or replace it with the alternate ASCII address.
If it is replaced, the replacement MUST be the ASCII-only address
specified with the ALT-ADDRESS parameter.[EAI-downgrading].

2.3. Extended Mailbox Address Syntax

RFC 2821, section 4.1.2, defines the syntax of a mailbox as

```
     Mailbox = Local-part "@" Domain

     Local-part = Dot-string / Quoted-string
           ; MAY be case-sensitive

     Dot-string = Atom *("." Atom)

     Atom = 1*atext

     Quoted-string = DQUOTE *qcontent DQUOTE

     Domain = (sub-domain 1*("." sub-domain)) / address-literal
     sub-domain = Let-dig [Ldh-str]
```

The key changes made by this specification are, informally, to

o Change the definition of "sub-domain" to permit either the
   definition above or a UTF-8 string representing a DNS label that
   is conformant with IDNA [RFC3490]. That label MUST NOT contain
   the characters "@" or ".", even though those characters can
   normally be inserted into a DNS label.
o Change the definition of "Atom" to permit either the definition
   above or a UTF-8 string. That string MUST NOT contain any of the
   ASCII characters (either graphics or controls) that are not

permitted in "atext"; it is otherwise unrestricted.

According to the description above, define the syntax of an
internationalized email mailbox with ABNF [RFC4234] as

         Mailbox = Local-part "@" Domain

         Local-part = Dot-string / Quoted-string
               ; MAY be case-sensitive

         Dot-string = Atom *("." Atom)

         Atom = 1*Ucharacter
         Ucharacter = atext / UTF8-2 / UTF8-3 / UTF8-4

         Quoted-string = DQUOTE *qcontent DQUOTE

         Domain = (sub-domain 1*("." sub-domain)) / address-literal
         sub-domain = ULet-dig [ULdh-str]

         ULet-dig = Let-dig / Non-ASCII

         ULdh-str = *(ALPHA / DIGIT / "-" / Non-ASCII) ULet-dig

         Non-ASCII = UTF8-2 / UTF8-3 / UTF8-4

   Where "atext", "qcontent" and "DQUOTE" are defined in [RFC2822],
   "Let-dig", "Ldh-str" and "address-literal" are defined in [RFC2821]
   and UTF8-2, UTF8-3 and UTF8-4 are defined in [RFC3629]. The value of
   "domain" SHOULD be verified with [RFC3490]; If failed, the email
   address with that domain can not be regarded as the valid email
   address.

2.4. The ALT-ADDRESS parameter

   If the UTF8SMTP extension is offered, the syntax of the SMTP MAIL and
   RCPT commands is extended to support the optional "ALT-ADDRESS"
   parameter.

   The "ALT-ADDRESS" requires an all-ASCII address.

   The ALT-ADDRESS parameter usage in the commands of "mail from" and
   "rcpt to" is defined according to the definition of mail-parameters
   in [RFC2821] below.

```
        MAIL FROM:<reverse-path> [SP <mail-parameters> ] <CRLF>
        RCPT TO:<forward-path> [ SP <rcpt-parameters> ] <CRLF>
        Mail-parameters = esmtp-param *(SP esmtp-param)
        Rcpt-parameters = esmtp-param *(SP esmtp-param)
        esmtp-param     = esmtp-keyword ["=" esmtp-value]
        esmtp-keyword   = (ALPHA / DIGIT) *(ALPHA / DIGIT / "-")
        esmtp-value     = 1*(%d33-60 / %d62-127)
              ; any CHAR excluding "=", SP, and control characters
        Reverse-path = Path
        Forward-path = Path
        Path = "<" [ A-d-l ":" ] Mailbox ">"
        A-d-l = At-domain *( "," A-d-l )
           ; Note that this form, the so-called "source route",
           ; MUST BE accepted, SHOULD NOT be generated, and SHOULD be
           ; ignored.
        At-domain = "@" domain
```

where the value of esmtp-keyword is "ALT-ADDRESS" and the value of
esmtp-value is all-ASCII email address, and where the domain and
Mailbox are defined at section 2.3 of this document.

The use of the ALT-ADDRESS is specified below: If some involved SMTP
servers can not support UTF8SMTP capability and if the sender has
already set the ALT-ADDRESS value, the client SMTP server will use
this address as the email address when the SMTP server does the
subsequent operations. If the ALT-ADDRESS value is not set by the
sender, the email must be bounced to the original sender. If the
email is bounced due to the incapability of supporting UTF8SMTP, the
relative server should issue the response error code "5.3.3" defined
in [RFC3463] which means that System is not capable of selected
features, permanent failure.

2.5. The Suggestion of the Value of the ALT-ADDRESS parameter

The "ALT-ADDRESS" requires an all-ASCII address. There are two
alternative ways to set ALT-ADDRESS value: one is set by the sender
using the all-ASCII address, the other is set using the transformed
email address.

Some may prefer transformed the non-ASCII address to the ASCII
Compatible Encoding(ACE) address as the value of the ALT-ADDRESS.
The big problem with applying an ACE to all local-parts is that the
sending or converting system doesn't know if there are some specific
data or instructions embedded in the address that the ACE process
would hide. Some SMTP servers may depend on these specific data or
instructions to do some operations while the local parts applied with
ACE will lose or hide these data or instructions. SMTP [RFC2821]

prohibits SMTP relays from converting local parts because the level
of SMTP relays' knowledge on the structure of local parts is assumed
to be zero. However, we can raise the knowledge level by supplying
additional information. Many human users' email addresses do not
have any embedded structure processed by the final delivery MTA. In
that case, the sender can specify that these email addresses are safe
to be converted in predefined way. The final delivery SMTP server
can revert the addresses even though they are as in all ASCII form.
Unless the MUA or the submission server clearly knows that the non-
ASCII address can be safely transformed into the all-ASCII address,
the non-ASCII address should not be transformed because transformed
email address may cause some potential problems.

This document suggests that the ALT-ADDRESS is set directly by the
sender; In default, the all-ASCII address can not be transformed.

2.6. Additional ESMTP Changes and Clarifications

The mail transport process involves addresses ("mailboxes") and
domain names in contexts in addition to the MAIL and RCPT commands
and extended alternatives to them. In general, the rule is that,
when RFC 2821 specifies a mailbox, this document expects UTF-8 to be
used for the entire string; when RFC 2821 specifies a domain name,
the name SHOULD be in punycode form if its raw form is non-ASCII.

The following subsections list and discuss all of the relevant cases.

Support and use of this extension requires support for 8BITMIME. It
means that 8BITMIME MUST be advertised by the UTF8SMTP capability
SMTP server.

2.6.1. The Initial SMTP Exchange

When an SMTP or ESMTP connection is opened, the server sends a
"banner" response consisting of the 220 reply code and some
information. The client then sends the EHLO command. Since the
client cannot know whether the server supports UTF8SMTP until after
it receives the response from EHLO, any domain names that appear in
this dialogue, or in responses to EHLO, MUST be in hostname form,
i.e., internationalized ones MUST be in punycode form.

2.6.2. Trace Fields

Internationalized domain names in Received fields MUST be transmitted
in the punycode form. Addresses in "for" clauses need further
examination and might be treated differently depending on [EAI-
utf8header]. The reasoning in the introductory portion of [EAI-
overview] strongly suggests that these addresses be in UTF-8 form,

rather than some specialized encoding.

2.6.3. Mailing List Question

How a mixture of traditional and internationalized addresses on a
mailing list will impact message flows, error reports, and delivery
notifications in all plausible combinations of UTF8SMTP capability
and un-capability servers is discussed and specified in the [EAI-
mailing list].

2.6.4. Message Header Label

The message header label MAY be used to identify and distinguish the
i18mail message from the normal message when SMTP messages are
transmitted on wire. This issue is discussed and specified in [EAI-
utf8header].

2.6.5. POP and IMAP

While SMTP mainly takes care of the transportation of messages and
the header fields on wire, POP essentially handles the retrieval of
mail objects from the server by a client. In order to use
internationalized user names based on i18mail for the retrieval of
messages from a mail server using the POP protocol, a new capability
SHOULD be introduced following the POP3 extension mechanism
[RFC2449]. This is discussed and specified in the [EAI-pop].

IMAP [RFC3501] uses the traditional user name which is based on
ASCII. IMAP SHOULD be updated to support the internationalized user
names based on i18mail for the retrieval of messages from a mail
server. This is discussed and specified in the [EAI-imap].

3. Potential problems

3.1. Impact to IRI

The mailto: schema in IRI [RFC3987] MAY need to be modified when EAI
is standardized.

3.2. Impact to RFC 2476 and many email related RFC

The EAI protocol will impact on many email related RFC such as
Message Submission [RFC2476] and SMTP Service Extension for DSNs
[RFC3461]. These protocol SHOULD be considered when implementing the
EAI protocol.

4. Implementation Advice

In the absence of this extension, SMTP clients and servers are
constrained to using only those addresses permitted by RFC 2821. The
local parts of those addresses MAY be made up of any ASCII
characters, although certain of them MUST be quoted as specified
there. It is notable in an internationalization context that there
is a long history on some systems of using overstruck ASCII
characters (a character, a backspace, and another character) within a
quoted string to approximate non-ASCII characters. This form of
internationalization SHOULD be phased out as this extension becomes
widely deployed but backward-compatibility considerations require
that it continue to be supported.

5. IANA Considerations

IANA is requested to add "UTF8SMTP" to the SMTP extensions registry
with the entry pointing to this specification for its definition.

6. Security considerations

See the extended security considerations discussion in [EAI-overview]

7. Acknowledgements

Much of the text in the initial version of this document was derived
or copied from [Klensin-emailaddr] with the permission of the author.
Significant comments and suggestions were received from Xiaodong LEE,
Nai-Wen Hsu, Yangwoo KO, Yoshiro YONEYA, and other members of the JET
team and were incorporated into the document. Special thanks to
those contributors for this version of document, those includes (but
not limited to) John C Klensin, Charles Lindsey, Dave Crocker, Harald
Tveit Alvestrand, Marcos Sanz, Chris Newman, Martin Duerst, Edmon
Chung, Tony Finch.

8. References

8.1. 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-downgrading]
YONEYA, Y., Ed. and K. Fujiwara, Ed., "Downgrading
mechanism for Internationalized eMail Address (IMA)",
draft-ietf-eai-downgrade-00 (work in progress),
October 2005.

[EAI-imap]
Resnick, P. and C. Newman, "Considerations for IMAP in
Conjunction with Email Address Internationalization",
draft-ietf-eai-imap-utf8-00 (work in progress), May 2006.

[EAI-mailing list]
Chung, E., "Mailing Lists and Internationalized Email
Addresses", June 2006.

Forthcoming

[EAI-overview]
Klensin, J. and Y. Ko, "Overview and Framework for
Internationalized Email", draft-ietf-eai-framework-01.txt
(work in progress), June 2006.

[EAI-pop]   Newman, C., "POP3 Support for UTF-8", June 2006, <http://
www.ietf.org/internet-drafts/draft-ietf-eai-pop-00.txt>.

[EAI-utf8header]
Yeh, J., "Transmission of Email Headers in UTF-8
Encoding", draft-ietf-eai-utf8headers-00.txt (work in
progress), June 2006.

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

[RFC1869]   Klensin, J., Freed, N., Rose, M., Stefferud, E., and D.
Crocker, "SMTP Service Extensions", STD 10, RFC 1869,
November 1995.

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

[RFC2449]   Gellens, R., Newman, C., and L. Lundblade, "POP3 Extension
Mechanism", RFC 2449, November 1998.

[RFC2476]   Gellens, R. and J. Klensin, "Message Submission",

RFC 2476, December 1998.

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

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

[RFC3454]   Hoffman, P. and M. Blanchet, "Preparation of
            Internationalized Strings ("stringprep")", RFC 3454,
            December 2002.

[RFC3461]   Moore, K., "Simple Mail Transfer Protocol (SMTP) Service
            Extension for Delivery Status Notifications (DSNs)",
            RFC 3461, January 2003.

[RFC3463]   Vaudreuil, G., "Enhanced Mail System Status Codes",
            RFC 3463, January 2003.

[RFC3490]   Faltstrom, P., Hoffman, P., and A. Costello,
            "Internationalizing Domain Names in Applications (IDNA)",
            RFC 3490, March 2003.

[RFC3492]   Costello, A., "Punycode: A Bootstring encoding of Unicode
            for Internationalized Domain Names in Applications
            (IDNA)", RFC 3492, March 2003.

[RFC3501]   Crispin, M., "INTERNET MESSAGE ACCESS PROTOCOL - VERSION
            4rev1", RFC 3501, March 2003.

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

[RFC3987]   Duerst, M. and M. Suignard, "Internationalized Resource
            Identifiers (IRIs)", RFC 3987, January 2005.

[RFC4234]   Crocker, D. and P. Overell, "Augmented BNF for Syntax
            Specifications: ABNF", RFC 4234, October 2005.

8.2. Informative References

[Klensin-emailaddr]
            Klensin, J., "Internationalization of Email Addresses",
            draft-klensin-emailaddr-i18n-03 (work in progress),
            July 2005.

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

Bodies", [RFC 2045](), November 1996.

Authors' Addresses

    Jiankang YAO (editor)
    CNNIC
    No.4 South 4th Street, Zhongguancun
    Beijing

    Phone: +86 10 58813007
    Email: yaojk@cnnic.cn

    Wei MAO (editor)
    CNNIC
    No.4 South 4th Street, Zhongguancun
    Beijing

    Phone: +86 10 58813055
    Email: mao@cnnic.cn