Internet Draft
draft-hoffman-utf8headers-00.txt
December 15, 2003
Expires in six months

Paul Hoffman
Internet Mail Consortium

SMTP Service Extensions or Transmission of Headers in UTF-8 Encoding

Status of this Memo

Abstract

Mailbox names often represent the names of human users. Many of these users throughout the world have names that are not normally represented by the users with just the ASCII repertoire of characters, and would therefore 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 how to represent all headers of email messages encoded in UTF-8.

1. Introduction

The format of email messages [MSGFMT] only allows ASCII characters in the headers 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 quoted-printable format [MIME3]. This specification describes a

change to the email message format, and to SMTP message transport, that allows non-ASCII characters throughout email headers. These changes affect SMTP clients, SMTP servers, and mail user agents (MUAs).

In this specification, the SMTP protocol [SMTP] is used to prevent the transmission of messages with UTF-8 [UTF8] headers to systems that cannot handle such messages. The new SMTP extension has the name "UTF-8-HEADERS".

Using this new SMTP extension prevents the introduction of such messages in message stores that might misrepresent or mangle such messages. It should be noted that using an ESMTP extension does not prevent transferring email messages with UTF-8 headers to other systems that use the email format for messages, such as in the POP and IMAP protocols. Those protocols will need to be changed in order to handle messages in message stores that have UTF-8 headers.

The dual motivations of this protocol are to allow UTF-8 everywhere in the headers and to not bounce any messages just because they originated with UTF-8 headers. Using this protocol, messages that originated with UTF-8 headers will only be bounced if an enabled SMTP client is speaking to an unenabled SMTP server and some of the UTF-8 headers cannot be downgraded to all-ASCII headers. This protocol describes how to downgrade all headers from UTF-8 to all-ASCII, but does not guarantee that such downgrading will always be successful.

Further, this protocol allows current users who have all-ASCII mailbox names to step up to UTF-8 headers easily. This means that users of this protocol should normally be able to communicate with other users of this protocol and with users who have not yet updated.

This protocol does not require the sender or recipient of mail to have mailbox names that do not include non-ASCII characters. For example, the protocol might still be used if just the subject header has non-ASCII characters, and the protocol must be used if other headers (particularly Received headers) contain non-ASCII characters.

1.1 Terminology

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

Unless otherwise noted, all terms used here are defined in RFC 2821 and RFC 2822.

In this document, an address is "all-ASCII" if every character in the address is in the ASCII character repertoire [ASCII]; an address is "non-ASCII" if any character is not in the ASCII character repertoire. Similarly, a header body is "all-ASCII" if every character in the body of the header is in the ASCII character repertoire; a header body is

"non-ASCII" if any character is not in the ASCII character repertoire.

This document is being discussed on the ietf-imaa mailing list. See <http://www.imc.org/ietf-imaa/> for information about subscribing and the list's archive.

2. Changes to MUAs and to the user's mail environment

For this protocol to work well (that is, for it not to bounce mail excessively when an enabled system encounters a non-enabled system), any mail sender who has non-ASCII characters in the addr-spec of their mailbox name SHOULD have a second mailbox whose addr-spec contains only ASCII characters. This second mailbox is used when a recipient of a message is not using this protocol; this is the "fallback address" for the sender.

Having two mailboxes is not an absolute requirement because some mail systems will not allow a user to be able to get mail from two addresses (the non-ASCII and all-ASCII addresses). If a user does have two mailboxes, they SHOULD both be on the same mail server (that is, they should both have the same host name in the user's address).

Having two mailboxes can lead to confusion for users if the MUA does not handle them well. MUAs that follow this specification SHOULD have options that would make it seem like two mailboxes are one. For example, if a user says "read my mail", the MUA SHOULD read from both the mailbox with the non-ASCII name and the mailbox with the all-ASCII name. Note that this feature might not be necessary: a terminating SMTP server might have combined all incoming mail for both addresses into a single mailbox. However, MUAs SHOULD NOT assume that combining by the SMTP server will always be the case.

2.1 Changes to MUA administrative interfaces

The administrative interface for MUAs that use this protocol MUST have method for a user to specify the name of their mailbox that contains non-ASCII characters, and MUST have a method for the user to specify the name of their mailbox that contains non-ASCII characters.

The MUA user interface SHOULD also allow users to specify the common name associated with the non-ASCII mailbox using non-ASCII characters; this common name MUST be encoded as UTF-8. The common name associated with the all-ASCII mailbox MUST only contain ASCII characters, although it can use a quoted-printable format to represent a different encoding; this encoding SHOULD be UTF-8.

MUAs are encouraged to cache address mappings that are specified in incoming mail. Given that mappings might change over time, these MUAs might over-write existing mappings with new ones, and might give the user a choice for the time-to-live for the

cached mapping.

2.2 Address-map headers

For every address in a message with a non-ASCII local-part, the mail initiator SHOULD create a mapping in a new header, called "Address-map:". A message SHOULD have one Address-map: header for every non-ASCII address for which the sender knows a map. The header is only for addresses that have a non-ASCII local-part in its addr-spec. It MUST NOT be used for addresses that have all-ASCII addr-specs, even if those addresses have UTF-8 domain names, and it MUST NOT be used if the local-part of the addr-spec is all-ASCII but the display-name or the comment is non-ASCII.

If the sender has an all-ASCII local-part associated with its non-ASCII mailbox, the sender's MUA MUST create an Address-map header for that association. If the sender knows (such as through caching incoming address maps or from an address book) the mapping for any recipient that has a non-ASCII mailbox name, the sending MUA SHOULD create an Address-map header for it.

Both addresses in the Address-map header are full addr-specs. The body of the Address-map header only contains addr-specs, never display-names or comments. The format of the Address-map header is:

```
   Address-map: <address-with-non-ASCII-LHS>,<downgrade-address>
```

The encoding for address-with-non-ASCII-LHS MUST be UTF-8; the encoding for downgrade-address MUST be ASCII. If the domain name in an internationalized domain name [IDNA], then it MUST be encoded in UTF-8 in the address-with-non-ASCII-LHS and MUST be encoded using IDNA in the downgrade-address.

Examples:

```
  Address-map: Jos<eacute>@example.com,jose@example.com

  Address-map: bj<oumlaut>n@r<aumlaut>ksm<oumlaut>rg<aring>s.se,
      bjorn-ascii@rksmrgs-5wao1o.se
```

Note that when receiving mail, the Address-map headers may be all in ASCII. This would be due to an intervening SMTP server or other agent downgrading the map. All-ASCII Address-map headers MUST be accepted.

2.3 Changes to MUA sending

Sending MUAs that follow this protocol MUST create all headers encoded in UTF-8. No other direct encodings are allowed. MUAs MAY continue to use quoted-printable text to specify some text in other encodings; however this is not recommended because it is likely that this will not interoperate well with MUAs that follow this specification.

3. Changes to SMTP

This protocol defines a new SMTP extension, UTF-8-HEADERS. (The formal definition is in the IANA Considerations section.)

3.1 UTF-8-HEADERS extension

If an SMTP server advertises the UTF-8-HEADERS extentension, an SMTP client that supports this protocol SHOULD send message headers as described in this document.

The terminal SMTP server is responsible for knowing whether or not the message store can handle UTF-8 headers. A terminal SMTP server MUST NOT advertise the UTF-8-HEADERS extension if the message store for which it is responsible cannot
handle UTF-8 headers.

If an SMTP client does not see the UTF-8-HEADERS extension advertised by an SMTP server, the SMTP client MUST downgrade the non-ASCII contents of all header bodies before continuing to send the message. The SMTP client SHOULD send the message with the downgraded header bodies as a normal message.
If any header body cannot be downgraded, the SMTP client MUST bounce the message with an error code of 558.

All UTF-8 headers bodies can be downgraded to being all-ASCII. However, any header body that contains a non-ASCII mailbox name might not be able to be downgraded if there is no Address-map header that gives a mapping for the downgrading.

3.2 Downgrading header bodies

This section defines how to downgrade header bodies. Note that downgrading MUST only be done if necessary. That is, downgrading MUST never be done on fields or bodies that are all-ASCII.

3.2.1 Mailboxes

Mailboxes appear in many standard headers, such as To:, From:, Sender:, Reply-to:, Cc:, Bcc:, Received:, and some of the Resent-: headers. Downgrading mailboxes is done as follows:

1) If necessary, convert the domain using IDNA.

2) If necessary,convert the local-parts using values from an Address-map: header in the message

3) If necessary,convert any display-name or comment using quoted-printable with UTF-8 encoding

3.3.2 Message-ids

Downgrading message-ids is done as follows

1) If necessary,convert the id-left using Base64

2) If necessary,convert the id-right using Base64

3.3.3 Informational headers

If necessary, downgrading the bodies of informational headers (Subject:, Comments:, and Keywords:) is done using quoted-printable with UTF-8 encoding.

3.3.4 Address-map headers

If necessary, the Address-map: header is downgraded using Base64 for local-parts, and IDNA for domain names.

For example:

```
Address-map: Jos<eacute>@example.com,jose@example.com
```

would be downgraded to:

```
Address-map: Sm9zw6k=@example.com,jose@example.com
```

As another example:

```
Address-map: bj<oumlaut>n@r<aumlaut>ksm<oumlaut>rg<aring>s.se,
    bjorn-ascii@rksmrgs-5wao1o.se
```

would be downgraded to:

```
Address-map: YmrDtnJu@rksmrgs-5wao1o.se,
    bjorn-ascii@rksmrgs-5wao1o.se
```

3.3 Things not changed from RFC 2822

Note that this protocol does change the definition of header field names. That is, only the bodies of headers are allowed to have non-ASCII characters; the rules in RFC 2822 for header names are not changed.

Similarly, this protocol does not change the date and time specification in RFC 2822.

3.4 Additional processing rules

In order to make mail retrieval easier, terminal SMTP servers SHOULD write messages addressed to either the UTF-8 address or the all-ASCII

address into the same mailbox. However, given that this is quite different than common practice today, the ramifications for doing this should be studied carefully before this is implemented.

Intermediate SMTP servers MAY change the values in the Address-map: header (such as to add one that is missing or to correct a mapping), but SHOULD only do so for domains local to the intermediate SMTP server.

Terminal SMTP servers MAY look into the headers of a message to determine whether they should upgrade a downgraded set of headers to UTF-8. This is easy to determine: if the Address-map: header contains only ASCII, it was downgraded earlier in the chain of SMTP server. Upgrading is particularly useful on bounce messages caused by bad mappings.

4. Security considerations

If a user has a non-ASCII mailbox address and a mapped all-ASCII mailbox address, a digital certificate that identifies that user SHOULD have both addresses in the identity. Having multiple email addresses as identities in a single certificate is already supported in PKIX and OpenPGP.

Internationalized local parts will cause mail addresses to become longer, and possibly make it harder to keep lines in a header under 78 characters. Lines that are longer than 78 characters (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.

5. IANA considerations

IANA will assign the UTF-8-HEADERS extension for ESMTP.

The UTF-8 headers extension is defined as follows:

(1) The name of the SMTP service extension is "UTF-8 headers".

(2) The EHLO keyword value associated with the extension is UTF-8-HEADERS.

(3) No parameter is used with the UTF-8-HEADERS EHLO keyword.

(4) No additional parameters are added to either the MAIL FROM or RCPT TO commands.

(5) No additional SMTP verbs are defined by this extension.

(6) This document specifies how support for the extension affects the behavior of a server and client SMTP.

6. References

6.1 Normative references

[ASCII] Cerf, V., "ASCII format for Network Interchange", RFC 20, October 1969.

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

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

[MIME3] Moore, K., "MIME (Multipurpose Internet Mail Extensions) Part Three: Message Header Extensions for Non-ASCII Text", RFC 2047, November 1996.

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

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

[UTF8] Yergeau, F. "UTF-8, a Transformation Format of ISO 10646", RFC 3629, November 2003.

7. Author's address

Paul Hoffman
Internet Mail Consortium
127 Segre Place
Santa Cruz, CA 95060 USA
phoffman@imc.org

A. Open issues

- POP and IMAP might be updated to allow one request to bring in two or more mailboxes; otherwise, users will have to do two separate requests.

- It might be good to have a protocol for determining mappings, but it is not defined here.