

Transition of Internet Mail from
Just-Send-8
to 8bit-SMTP/MIME

Status of this Memo

This RFC provides information for the Internet community. It does not specify an Internet standard. Distribution of this memo is unlimited.

Abstract

Protocols for extending SMTP to pass 8bit characters have been defined [3] [4]. These protocols require that messages transported by the extended SMTP are to be encoded in MIME [1] [2]. Before work began on these protocols, several SMTP implementations adopted ad-hoc mechanisms for sending 8bit data. It is desirable for the extended SMTP environment and these ad hoc mechanisms interoperate. This document outlines the problems in this environment and an approach to minimizing the cost of transition from current usage of non-MIME 8bit messages to MIME.

1. Terminology

RFC 821 defines a 7bit transport. A transport agent which does not clear the high order bit upon receipt of octets with this bit set in SMTP messages is called 8 bit transparent in this document. An implementation of the general SMTP Extensions document [3] and the 8bit extensions protocol [4] which passes MIME messages using all 8 bits of an octet is called 8bit ESMTP. An implementation of extended SMTP which does not accept 8bit characters is called 7bit ESMTP. A gateway is defined to be a transport agent with User Agent authority to alter or convert the content of a message.

2. The Problem

SMTP as defined in RFC 821 limits the sending of Internet Mail to US-ASCII [5] characters. As the Internet has grown to include non-English correspondents, the need to communicate with character sets other than US-ASCII has prompted many vendors and users to extend SMTP or RFC 822 to use non-ASCII character sets. Common approaches are to send 7 bit national variant ISO 646 character sets over current RFC822/SMTP, to extend SMTP and RFC822 to use 8bit ISO 8859

character sets, and to use proprietary PC character sets.

A third approach is used for Japanese mail. Japanese characters are represented by pairs of octets with the high order bit cleared. Switching between 14 bit character sets and 7 bit character sets is indicated within the message by ISO 2022 escape sequences.

So long as these implementations can communicate without intermediate transformations and have a loose private agreement on the use of a specific character set without tagging, basic mail service can be provided.

In the transition to the negotiated 8bit ESMTMP/MIME environment, it is important that mail sent by a currently non-conforming user can be read by another non-conforming user. This existing functionality is reduced by conversion from 8bit text to text encoded in unreadable Base-64 or "garbled" text encoded in quoted printable.

There are several interesting non-interoperable cases that currently exist in non US-ASCII mail and several new ones likely to emerge in a transition to 8bit/MIME. Below is a listing of the transition-to-mime cases. Only solutions to (4) in the context of a translating gateway are discussed in this memo.

| | \ Receiver | | |
|------------------|------------|------------------|------------|
| Sender \ | 7bit only | 8bit transparent | MIME/ESMTP |
| 7bit only | (1) | (1) | (1) |
| 8bit transparent | (2) | (3) | (4) |
| MIME/ESMTP | (5) | (5) | (6) |

(1) 7Bit non-MIME sender to 7bit, MIME, or 8bit transparent receiver

This will continue to work unchanged with nationally variant ISO 646 or ISO 2022 character set shifting if an external "out of band" agreement exists between the sender and the receiver. A 7bit to 8bit/ESMTP gateway need not alter the content of this message.

(2) 8bit sender to 7bit non-MIME receiver

The receiver will receive bit-stripped mail which results in the mis-interpretation of the data and the wrong character being displayed or printed. Mail sent using languages where most

characters are in the US-ASCII subset of ISO 8859 may be somewhat readable.

(3) 8bit transparent sender to 8bit transparent receiver

Will work if an external agreement "out of band" to use a particular character set without tagging exists between the sender and the receiver.

(4) 8bit transparent sender to MIME/ESMTP conformant receiver

Will work if a reasonable upgrade path is provided via gateways, the indicated character set tag inserted by the gateway is correct and the receiver supports the character set chosen by the sender. This case is the focus of this memo.

(5) MIME/ESMTP sender to non-MIME 7bit receiver

Because the ESMTP/MIME sender cannot know if the receiver will understand 8bits, the sender will encode the text into base-64 or quoted-printable which may be considered "garbled" by the receiver. To provide a useful downgrade path the gateway must have some knowledge about the capabilities of the receiver. When the character set can be clearly identified, techniques like the mnemonic MNEM encoding described in RFC 1345 may be helpful in this case.

(6) MIME/ESMTP sender to MIME/ESMTP receiver

Interoperability will be attained provided the receiver supports the character set chosen by the sender.

3. Upgrade Path from 8bit Transparent to ESMTP/MIME

A gateway which has been upgraded to support Extended SMTP may upgrade an 8bit message received to MIME. This is consistent with the requirement that all 8bit mail sent by ESMTP be encoded in MIME. The upgrade should be done using the best available information.

A site may "upgrade" to MIME en-masse by implementing MIME conversion for all messages leaving the site. For text messages, the body can be converted by adding a MIME-version header and a Content-Type: Text/Plain with the character set in use in the site, provided the site uses a single character set.

An appropriate Content-Transfer-Encoding header line must be added to indicate any encoding that may be necessary.

Example:

```
MIME-Version: 1.0
Content-Type: Text/Plain; Charset = ISO-8859-1
Content-Transfer-Encoding: 8bit
```

If no information about the character set in use is available, the gateway should upgrade the content by using the character set "unknown-8bit". The unknown-8bit value of the charset parameter indicates only that no reliable information about the character set(s) used in the message was available.

If a message body has been upgraded to MIME, the RFC 822 headers containing non US-ASCII characters must be upgraded to conform with the header encoding rules of RFC1342. A gateway should recode all unstructured header fields as well as RFC 822 "comment"s and "phrase"s according to the rules of RFC 1342. There is no equivalent in RFC 1342 to the "8bit" Content-Transfer-Encoding value for message bodies so all 8bit header text must be transformed according to either the "B" or the "Q" encoding method. For ISO 8859 character sets, the "Q" encoding will generally result in somewhat readable headers.

Trace information should be added to the document with a convert clause: "rfc822-to-8bit", "rfc822-to-base-64" or "rfc822-to-quoted-printable" e.g.,

```
Received: from dbc.mtview.ca.us by dbc.mtview.ca.us
        convert rfc822-to-8bit; Tue, 01 Sep 1992 01:18:00 -0700
```

Appendix - The "unknown-8bit" Character Set

This section defines a "charset" parameter, for use in a MIME Content-Type field.

A special purpose character set called "unknown-8bit" is defined to be an unknown 8bit character set, encoded into a sequence of octets. It can be used as a label for any character set from any language, using any encoding. It must not be further defined.

The use of this token in a "charset=" field of a message indicates that nothing is known about the character set used. This marker is intended for use by non-MIME to MIME gateways; specifically in those which translate from SMTP to 8bit ESMTTP/MIME.

This character set is not intended to be used by mail composers. It is assumed that the mail composer knows the character set in use and will mark it with a character set value as specified in [1], as

amended by current Assigned Numbers documents [6].

The use of the "unknown-8bit" label is intended only by mail gateway agents which cannot determine via out-of-band information the intended character set.

The interpretation of the "unknown-8bit" is up to the mail reader. It is assumed that in many cases the human user will be able to interpret the information and choose an appropriate character set or pre-processor.

Acknowledgements

This document originated as a hallway conversation between Ned Freed, Neil Katin, and the author. Substantive input was received from Jonathan Laventhol, Craig Everhart, Olle Jarnefors, and Olafur Gudmundsson. The document was refined with the input of many participants in the IETF SMTP Extensions Working Group.

References

- [1] Borenstein, N., and N. Freed, "Multipurpose Internet Mail Extensions", RFC 1341, Bellcore, Innosoft, June 1992.
- [2] Moore, K., "Representation of Non-ASCII Text in Internet Message Headers", RFC 1342, University of Tennessee, June 1992.
- [3] Klensin, J., WG Chair, Freed, N., Editor, Rose, M., Stefferud, E., and D. Crocker, "SMTP Service Extensions" RFC 1425, United Nations University, Innosoft International, Inc., Dover Beach Consulting, Inc., Network Management Associates, Inc., The Branch Office, February 1993.
- [4] Klensin, J., WG Chair, Freed, N., Editor, Rose, M., Stefferud, E., and D. Crocker, "SMTP Service Extensions for 8bit MIMETransport", RFC 1426, United Nations University, Innosoft International, Inc., Dover Beach Consulting, Inc., Network Management Associates, Inc., The Branch Office, February 1993.
- [5] Coded Character Set--7-Bit American Standard Code for Information Interchange, ANSI X3.4-1986.
- [6] Reynolds, J., and J. Postel, "Assigned Numbers", STD 2, RFC 1340, USC/Information Sciences Institute, July 1992.

Security Considerations

Security issues are not discussed in this memo.

Author's Address

Greg Vaudreuil
Corporation for National Research Initiatives
1895 Preston White Drive, Suite 100
Reston, VA 22091 USA

Phone: (703) 620-8990
EMail: GVaudre@CNRI.Reston.VA.US