Transmission of Email Headers in UTF-8 Encoding
draft-yeh-ima-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 March 31, 2006.

Copyright Notice

Copyright (C) The Internet Society (2005).

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 international characters in envelope addresses. It also requires being able to express those addresses and information based on them in mail headers. This document specifies the use of Unicode encoded in UTF-8, rather than ASCII, as the base form for Internet email headers. 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
   1.2. Background and History .................................... 3
   1.3. Terminology .................................................. 4
2. Changes to MUAs and to the user’s mail environment .............. 4
   2.1. Changes to MUA sending ...................................... 4
   2.2. Changes to MUA receiving .................................... 5
3. Changes to SMTP Servers and Clients ............................... 5
   3.1. Impact on Message Headers ................................. 5
   3.2. Things not changed from RFC 2822 ......................... 6
   3.3. Additional processing rules ............................... 6
4. Security Considerations ............................................. 6
5. IANA considerations .................................................. 6
6. Acknowledgements ................................................... 6
7. References ............................................................ 7
   7.1. Normative References ........................................ 7
   7.2. Informative References ..................................... 7
Author’s Address ......................................................... 9

Intellectual Property and Copyright Statements ..................... 10
1. Introduction

1.1. Role of this specification

Full internationalization of electronic mail requires several capabilities:

- The capability to transmit non-ASCII content, provided for as part of the basic MIME specification [RFC2045], [RFC2046].
- The capability to encode selected information in specific header fields, provided for as another part of the MIME specification [RFC2047].
- The capability to use international characters in envelope addresses, discussed in [IMA-overview] and specified in [IMA-SMTP-extension]. And, finally,
- The capability to express those addresses, and information related to and based on them, in mail headers, 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 headers. This form is permitted in transmission, if and only if authorized by the SMTP extension specified in [IMA-SMTP-extension].

1.2. 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 messages.

The traditional format of email messages [RFC2822] 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 [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 [IMA-overview] and [IMA-SMTP-extension], and that allows non-ASCII characters throughout email headers. These changes affect SMTP clients, SMTP servers, and mail user agents (MUAs).

As specified in [IMA-SMTP-extension], an SMTP protocol extension [RFC2821] is used to prevent the transmission of messages with UTF-8
headers 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 headers 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 headers.

The objective for this protocol is to allow UTF-8 everywhere in the headers. Issues about how to handle messages that contain UTF-8 headers but are proposed to be delivered to systems that have not been upgraded to support this capability are discussed elsewhere, particularly in [IMA-downgrading].

This protocol is workable even if IMA mailbox names 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 headers (particularly trace headers such as "Received:") contain non-ASCII characters.

1.3. Terminology

In this document, headers 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 [IMA-overview].

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

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

2.1. 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.

2.2. Changes to MUA receiving

Receiving MUAs that follow this protocol MUST able to handle email headers encoded in UTF-8. Which means that the email fetching protocol such as POP3 or IMAP MAY need to be updated.

3. Changes to SMTP Servers and Clients

The use of UTF-8 headers is dependent on the use of an SMTP extension named "IMA".

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

3.1. Impact on Message Headers

If an SMTP server advertises the IMA extension, an SMTP client that supports this protocol SHOULD send message headers as described in this document.

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

If an SMTP client see the IMA extension advertised by an SMTP server, the SMTP client MUST send all header message in UTF-8. However, the Message-ID is the unique identifier of a single email. In order to maintain the identity, message identifiers of the Message-ID fields MUST be created in all ASCII. Also when In-Reply-To or Reference are presented in email header, the Message-ID in these header fields MUST be created in all ASCII.

If an SMTP client does not see the IMA extension advertised by an SMTP server, the SMTP client MAY either

- Downgrade the non-ASCII contents of all header bodies before continuing to send the message, as described in [IMA-downgrading]. The SMTP client SHOULD send the message with the downgraded header bodies as a normal message.
- Reject the message with a reply code of 558. If any header body cannot be downgraded, this second option MUST be chosen.
3.2. 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.

3.3. Additional processing rules

In order to make mail retrieval easier, final delivery SMTP servers SHOULD write messages addressed to either the non-ASCII 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.

4. Security Considerations

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

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 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

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

6. 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 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.

7. References

7.1. Normative References


ANSI X3.4-1968 has been replaced by newer versions with slight modifications, but the 1968 version remains definitive for the Internet.


7.2. Informative References

draft-hoffman-utf8headers-00.txt (work in progress), December 2003.


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
Intellectual Property Statement

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.

Disclaimer of Validity

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 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.

Copyright Statement

Copyright (C) The Internet Society (2005). 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.

Acknowledgment

Funding for the RFC Editor function is currently provided by the Internet Society.