Internet Engineering Task Force (IETF)
Request for Comments: 6532
Obsoletes: 5335
Updates: 2045
Category: Standards Track
ISSN: 2070-1721

A. Yang
TWNIC
S. Steele
Microsoft
N. Freed
Oracle

February 2012

Internationalized Email Headers

Abstract

Internet mail was originally limited to 7-bit ASCII. MIME added support for the use of 8-bit character sets in body parts, and also defined an encoded-word construct so other character sets could be used in certain header field values. However, full internationalization of electronic mail requires additional enhancements to allow the use of Unicode, including characters outside the ASCII repertoire, in mail addresses as well as direct use of Unicode in header fields like "From:", "To:", and "Subject:", without requiring the use of complex encoded-word constructs. This document specifies an enhancement to the Internet Message Format and to MIME that allows use of Unicode in mail addresses and most header field content.

This specification updates Section 6.4 of RFC 2045 to eliminate the restriction prohibiting the use of non-identity content-transfer-encodings on subtypes of "message/".

Status of This Memo

This is an Internet Standards Track document.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Further information on Internet Standards is available in Section 2 of RFC 5741.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at http://www.rfc-editor.org/info/rfc6532.

Table of Contents

1. Introduction

Internet mail distinguishes a message from its transport and further divides a message between a header and a body [RFC5322]. Internet mail header field values contain a variety of strings that are intended to be user-visible. The range of supported characters for these strings was originally limited to [ASCII] in 7-bit form. MIME [RFC2045] [RFC2046] [RFC2047] provides the ability to use additional character sets, but this support is limited to body part data and to special encoded-word constructs that were only allowed in a limited number of places in header field values.

Globalization of the Internet requires support of the much larger set of characters provided by Unicode [RFC5198] in both mail addresses and most header field values. Additionally, complex encoding schemes like encoded-words introduce inefficiencies as well as significant opportunities for processing errors. And finally, native support for the UTF-8 charset is now available on most systems. Hence, it is strongly desirable for Internet mail to support UTF-8 [RFC3629] directly.

This document specifies an enhancement to the Internet Message Format [RFC5322] and to MIME that permits the direct use of UTF-8, rather than only ASCII, in header field values, including mail addresses. A new media type, message/global, is defined for messages that use this extended format. This specification also lifts the MIME restriction on having non-identity content-transfer-encodings on any subtype of the message top-level type so that message/global parts can be safely transmitted across existing mail infrastructure.

This specification is based on a model of native, end-to-end support for UTF-8, which depends on having an "8-bit-clean" environment assured by the transport system. Support for carriage across legacy, 7-bit infrastructure and for processing by 7-bit receivers requires additional mechanisms that are not provided by these specifications.

This specification is a revision of and replacement for [RFC5335]. Section 6 of [RFC6530] describes the change in approach between this specification and the previous version.

2. Terminology Used in This Specification

A plain ASCII string is fully compatible with [RFC5321] and [RFC5322]. In this document, non-ASCII strings are UTF-8 strings if they are in header field values that contain at least one <UTF8-non-ascii> (see Section 3.1).

Unless otherwise noted, all terms used here are defined in [RFC5321], [RFC5322], [RFC6530], or [RFC6531].

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

The term "8-bit" means octets are present in the data with values above 0x7F.

3. Changes to Message Header Fields

To permit non-ASCII Unicode characters in field values, the header definition in [RFC5322] is extended to support the new format. The following sections specify the necessary changes to RFC 5322's ABNF.

The syntax rules not mentioned below remain defined as in [RFC5322].

Note that this protocol does not change rules in RFC 5322 for defining header field names. The bodies of header fields are allowed to contain Unicode characters, but the header field names themselves must consist of ASCII characters only.

Also note that messages in this format require the use of the SMTPUTF8 extension [RFC6531] to be transferred via SMTP.

3.1. UTF-8 Syntax and Normalization

UTF-8 characters can be defined in terms of octets using the following ABNF [RFC5234], taken from [RFC3629]:

```
UTF8-non-ascii  =   UTF8-2 / UTF8-3 / UTF8-4

UTF8-2          =   <Defined in Section 4 of RFC3629>

UTF8-3          =   <Defined in Section 4 of RFC3629>

UTF8-4          =   <Defined in Section 4 of RFC3629>
```

See [RFC5198] for a discussion of Unicode normalization; normalization form NFC [UNF] SHOULD be used. Actually, if one is going to do internationalization properly, one of the most often cited goals is to permit people to spell their names correctly. Since many mailbox local parts reflect personal names, that principle applies to mailboxes as well. The NFKC normalization form [UNF] SHOULD NOT be used because it may lose information that is needed to correctly spell some names in some unusual circumstances.

3.2. Syntax Extensions to RFC 5322

The following rules extend the ABNF syntax defined in [RFC5322] and [RFC5234] in order to allow UTF-8 content.

```
VCHAR   =/  UTF8-non-ascii

ctext   =/  UTF8-non-ascii

atext   =/  UTF8-non-ascii

qtext   =/  UTF8-non-ascii

text    =/  UTF8-non-ascii
              ; note that this upgrades the body to UTF-8

dtext   =/  UTF8-non-ascii
```

The preceding changes mean that the following constructs now allow UTF-8:

1. Unstructured text, used in header fields like "Subject:" or "Content-description:".

2. Any construct that uses atoms, including but not limited to the local parts of addresses and Message-IDs. This includes addresses in the "for" clauses of "Received:" header fields.

3. Quoted strings.

4. Domains.

Note that header field names are not on this list; these are still restricted to ASCII.

3.3. Use of 8-bit UTF-8 in Message-IDs

Implementers of Message-ID generation algorithms MAY prefer to restrain their output to ASCII since that has some advantages, such as when constructing "In-reply-to:" and "References:" header fields in mailing-list threads where some senders use internationalized addresses and others do not.

3.4. Effects on Line Length Limits

Section 2.1.1 of [RFC5322] limits lines to 998 characters and recommends that the lines be restricted to only 78 characters. This specification changes the former limit to 998 octets. (Note that, in

ASCII, octets and characters are effectively the same, but this is not true in UTF-8.) The 78-character limit remains defined in terms of characters, not octets, since it is intended to address display width issues, not line-length issues.

3.5. Changes to MIME Message Type Encoding Restrictions

This specification updates Section 6.4 of [RFC2045]. [RFC2045] prohibits applying a content-transfer-encoding to any subtypes of "message/". This specification relaxes that rule -- it allows newly defined MIME types to permit content-transfer-encoding, and it allows content-transfer-encoding for message/global (see Section 3.7).

Background: Normally, transfer of message/global will be done in 8-bit-clean channels, and body parts will have "identity" encodings, that is, no decoding is necessary.

But in the case where a message containing a message/global is downgraded from 8-bit to 7-bit as described in [RFC6152], an encoding might have to be applied to the message. If the message travels multiple times between a 7-bit environment and an environment implementing these extensions, multiple levels of encoding may occur. This is expected to be rarely seen in practice, and the potential complexity of other ways of dealing with the issue is thought to be larger than the complexity of allowing nested encodings where necessary.

3.6. Use of MIME Encoded-Words

The MIME encoded-words facility [RFC2047] provides the ability to place non-ASCII text, but only in a subset of the places allowed by this extension. Additionally, encoded-words are substantially more complex since they allow the use of arbitrary charsets. Accordingly, encoded-words SHOULD NOT be used when generating header fields for messages employing this extension. Agents MAY, when incorporating material from another message, convert encoded-word use to direct use of UTF-8.

Note that care must be taken when decoding encoded-words because the results after replacing an encoded-word with its decoded equivalent in UTF-8 may be syntactically invalid. Processors that elect to decode encoded-words MUST NOT generate syntactically invalid fields.

3.7. The message/global Media Type

Internationalized messages in this format MUST only be transmitted as authorized by [RFC6531] or within a non-SMTP environment that supports these messages. A message is a "message/global message" if:

- it contains 8-bit UTF-8 header values as specified in this document, or
- it contains 8-bit UTF-8 values in the header fields of body parts.

The content of a message/global part is otherwise identical to that of a message/rfc822 part.

If an object of this type is sent to a 7-bit-only system, it MUST have an appropriate content-transfer-encoding applied. (Note that a system compliant with MIME that doesn't recognize message/global is supposed to treat it as "application/octet-stream" as described in Section 5.2.4 of [RFC2046].)

The registration is as follows:

Type name: message

Subtype name: global

Required parameters: none

Optional parameters: none

Encoding considerations: Any content-transfer-encoding is permitted. The 8-bit or binary content-transfer-encodings are recommended where permitted.

Security considerations: See Section 4.

Interoperability considerations: This media type provides functionality similar to the message/rfc822 content type for email messages with internationalized email headers. When there is a need to embed or return such content in another message, there is generally an option to use this media type and leave the content unchanged or down-convert the content to message/rfc822. Each of these choices will interoperate with the installed base, but with different properties. Systems unaware of internationalized headers will typically treat a message/global body part as an unknown attachment, while they will understand the structure of a message/rfc822. However, systems that understand message/global

will provide functionality superior to the result of a down-conversion to message/rfc822. The most interoperable choice depends on the deployed software.

Published specification: RFC 6532

Applications that use this media type: SMTP servers and email clients that support multipart/report generation or parsing. Email clients that forward messages with internationalized headers as attachments.

Additional information:

Magic number(s): none

File extension(s): The extension ".u8msg" is suggested.

Macintosh file type code(s): A uniform type identifier (UTI) of "public.utf8-email-message" is suggested. This conforms to "public.message" and "public.composite-content", but does not necessarily conform to "public.utf8-plain-text".

Person & email address to contact for further information: See the Authors' Addresses section of this document.

Intended usage: COMMON

Restrictions on usage: This is a structured media type that embeds other MIME media types. An 8-bit or binary content-transfer-encoding SHOULD be used unless this media type is sent over a 7-bit-only transport.

Author: See the Authors' Addresses section of this document.

Change controller: IETF Standards Process

4. Security Considerations

Because UTF-8 often requires several octets to encode a single character, internationalization may cause header field values (in general) and mail addresses (in particular) to become longer. As specified in [RFC5322], each line of characters MUST be no more than 998 octets, excluding the CRLF. On the other hand, MDA (Mail Delivery Agent) processes that parse, store, or handle email addresses or local parts must take extra care not to overflow buffers, truncate addresses, or exceed storage allotments. Also, they must take care, when comparing, to use the entire lengths of the addresses.

There are lots of ways to use UTF-8 to represent something equivalent or similar to a particular displayed character or group of characters; see the security considerations in [RFC3629] for details on the problems this can cause. The normalization process described in Section 3.1 is recommended to minimize these issues.

The security impact of UTF-8 headers on email signature systems such as Domain Keys Identified Mail (DKIM), S/MIME, and OpenPGP is discussed in Section 14 of [RFC6530].

If a user has a non-ASCII mailbox address and an ASCII mailbox address, a digital certificate that identifies that user might have both addresses in the identity. Having multiple email addresses as identities in a single certificate is already supported in PKIX (Public Key Infrastructure using X.509) [RFC5280] and OpenPGP [RFC3156], but there may be user-interface issues associated with the introduction of UTF-8 into addresses in this context.

5. IANA Considerations

IANA has updated the registration of the message/global MIME type using the registration form contained in Section 3.7.

6. Acknowledgements

This document incorporates many ideas first described in a draft document by Paul Hoffman, although many details have changed from that earlier work.

The authors especially thank Jeff Yeh for his efforts and contributions on editing previous versions.

Most of the content of this document was provided by John C Klensin and Dave Crocker. Significant comments and suggestions were received from Martin Duerst, Julien Elie, Arnt Gulbrandsen, Kristin Hubner, Kari Hurtta, Yangwoo Ko, Charles H. Lindsey, Alexey Melnikov, Chris Newman, Pete Resnick, Yoshiro Yoneya, and additional members of the Joint Engineering Team (JET) and were incorporated into the document. The authors wish to sincerely thank them all for their contributions.

7. References

7.1. Normative References

[ASCII] "Coded Character Set -- 7-bit American Standard Code for Information Interchange", ANSI X3.4, 1986.

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

[RFC3629] Yergeau, F., "UTF-8, a transformation format of ISO 10646", STD 63, RFC 3629, November 2003.

[RFC5198] Klensin, J. and M. Padlipsky, "Unicode Format for Network Interchange", RFC 5198, March 2008.

[RFC5234] Crocker, D. and P. Overell, "Augmented BNF for Syntax Specifications: ABNF", STD 68, RFC 5234, January 2008.

[RFC5321] Klensin, J., "Simple Mail Transfer Protocol", RFC 5321, October 2008.

[RFC5322] Resnick, P., Ed., "Internet Message Format", RFC 5322, October 2008.

[RFC6530] Klensin, J. and Y. Ko, "Overview and Framework for Internationalized Email", RFC 6530, February 2012.

[RFC6531] Yao, J. and W. Mao, "SMTP Extension for Internationalized Email", RFC 6531, February 2012.

[UNF] Davis, M. and K. Whistler, "Unicode Standard Annex #15: Unicode Normalization Forms", September 2010, <http://www.unicode.org/reports/tr15/>.

7.2. Informative References

[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.

[RFC3156] Elkins, M., Del Torto, D., Levien, R., and T. Roessler, "MIME Security with OpenPGP", RFC 3156, August 2001.

[RFC5280] Cooper, D., Santesson, S., Farrell, S., Boeyen, S., Housley, R., and W. Polk, "Internet X.509 Public Key Infrastructure Certificate and Certificate Revocation List (CRL) Profile", RFC 5280, May 2008.

[RFC5335] Yang, A., "Internationalized Email Headers", RFC 5335, September 2008.

[RFC6152] Klensin, J., Freed, N., Rose, M., and D. Crocker, "SMTP Service Extension for 8-bit MIME Transport", STD 71, RFC 6152, March 2011.

Authors' Addresses

Abel Yang
TWNIC
4F-2, No. 9, Sec 2, Roosevelt Rd.
Taipei 100
Taiwan

Phone: +886 2 23411313 ext 505
EMail: abelyang@twnic.net.tw

Shawn Steele
Microsoft

EMail: Shawn.Steele@microsoft.com

Ned Freed
Oracle
800 Royal Oaks
Monrovia, CA 91016-6347
USA

EMail: ned+ietf@mrochek.com