Network Working Group
Request for Comments: 1641
Category: Experimental

D. Goldsmith
M. Davis
Taligent, Inc.
July 1994

Using Unicode with MIME

Status of this Memo

This memo defines an Experimental Protocol for the Internet community. This memo does not specify an Internet standard of any kind. Distribution of this memo is unlimited.

Abstract

The Unicode Standard, version 1.1, and ISO/IEC 10646-1:1993(E) jointly define a 16 bit character set (hereafter referred to as Unicode) which encompasses most of the world's writing systems. However, Internet mail (STD 11, RFC 822) currently supports only 7-bit US ASCII as a character set. MIME (RFC 1521 and RFC 1522) extends Internet mail to support different media types and character sets, and thus could support Unicode in mail messages. MIME neither defines Unicode as a permitted character set nor specifies how it would be encoded, although it does provide for the registration of additional character sets over time.

This document specifies the usage of Unicode within MIME.

Motivation

Since Unicode is starting to see widespread commercial adoption, users will want a way to transmit information in this character set in mail messages and other Internet media. Since MIME was expressly designed to allow such extensions and is on the standards track for the Internet, it is the most appropriate means for encoding Unicode. RFC 1521 and RFC 1522 do not define Unicode as an allowed character set, but allow registration of additional character sets.

In addition to allowing use of Unicode within MIME bodies, another goal is to specify a way of using Unicode that allows text which consists largely, but not entirely, of US-ASCII characters to be represented in a way that can be read by mail clients who do not understand Unicode. This is in keeping with the philosophy of MIME. Such an encoding is described in another document, "UTF-7: A Mail Safe Transformation Format of Unicode" [UTF-7].

Overview

Several ways of using Unicode are possible. This document specifies both guidelines for use of Unicode within MIME, and a specific usage. The usage specified in this document is a straightforward use of Unicode as specified in "The Unicode Standard, Version 1.1".

This encoding is intended for situations where sender and recipient do not want to do a lot of processing, when the text does not consist primarily of characters from the US-ASCII character set, or when sender and receiver are known in advance to support Unicode.

Another encoding is intended for situations where the text consists primarily of US-ASCII, with occasional characters from other parts of Unicode. This encoding allows the US-ASCII portion to be read by all recipients without having to support Unicode. This encoding is specified in another document, "UTF-7: A Mail Safe Transformation Format of Unicode" [UTF-7].

Finally, in keeping with the principles set forth in RFC 1521, text which can be represented using the US-ASCII or ISO-8859-x character sets should be so represented where possible, for maximum interoperability.

Definitions

The definition of character set Unicode:

The 16 bit character set Unicode is defined by "The Unicode Standard, Version 1.1". This character set is identical with the character repertoire and coding of the international standard ISO/IEC 10646-1:1993(E); Coded Representation Form=UCS-2; Subset=300; Implementation Level=3.

Note. Unicode 1.1 further specifies the use and interaction of these character codes beyond the ISO standard. However, any valid 10646 BMP (Basic Multilingual Plane) sequence is a valid Unicode sequence, and *vice versa*; Unicode supplies interpretations of sequences on which the ISO standard is silent as to interpretation.

This character set is encoded as sequences of octets, two per 16-bit character, with the most significant octet first. Text with an odd number of octets is ill-formed.

Rationale. ISO/IEC 10646-1:1993(E) specifies that when characters in the UCS-2 form are serialized as octets, that the most significant octet appear first. This is also in keeping with common network practice of choosing a canonical format for transmission.

General Specification of Unicode Character Sets Within MIME

The Unicode Standard is currently at version 1.1. Although new versions should be compatible with old implementations if an implementation is compliant with the standard, some implementations may choose to check the version of the character set that is being used. In order to allow some implementations to check the version number and allow others to ignore it, all registrations of Unicode variants and versions for MIME usage should have MIME charset names which conform to one of the two following patterns:

```
UNICODE-major-minor
UNICODE-major-minor-variant
```

Where major and minor are strings of decimal digits (0 through 9) specifying the major and minor version number of the Unicode standard to which the text in question conforms. In the interests of interoperability, the lowest version number compatible with the text should be used. The lowest acceptable version number is UNICODE-1-1, corresponding to “The Unicode Standard, Version 1.1”. The optional trailing string “variant” describes the particular transformation format of Unicode that the registration describes; its content is up to the particular registration. If there is no trailing variant string, the charset name refers to the basic two octet form of Unicode as described in “The Unicode Standard”.

Example. A hypothetical charset which referred to the UTF-8 transformation format of Unicode/10646 (also known as UTF-2 or UTF-FSS) might be named UNICODE-1-1-UTF-8.

Encoding Character Set Unicode Within MIME

Character set Unicode uses 16 bit characters, and therefore would normally be used with the Binary or Base64 content transfer encodings of MIME. In header fields, it would normally be used with the B content transfer encoding. The MIME character set identifier is UNICODE-1-1.

Example. Here is a text portion of a MIME message containing the Japanese word “nihongo” (hexadecimal 65E5,672C,8A9E) written in Han characters.

```
Content-Type: text/plain; charset=UNICODE-1-1
Content-Transfer-Encoding: base64

ZeVnLIqe
```

Example. Here is a text portion of a MIME message containing the Unicode sequence “A<NOT IDENTICAL TO><ALPHA>.” (hexadecimal 0041,2262,0391,002E)

```
Content-Type: text/plain; charset=UNICODE-1-1
Content-Transfer-Encoding: base64

AEEiYgORAC4=
```

Acknowledgements

Many thanks to the following people for their contributions, comments, and suggestions. If we have omitted anyone it was through oversight and not intentionally.

Glenn Adams
Harald T. Alvestrand
Nathaniel Borenstein
Lee Collins
Jim Conklin
Dave Crocker
Steve Dorner
Dana S. Emery
Ned Freed
John H. Jenkins
John C. Klensin
Valdis Kletnieks
Keith Moore
Masataka Ohta
Einar Stefferud

Security Considerations

Security issues are not discussed in this memo.

References

[UNICODE 1.1] "The Unicode Standard, Version 1.1": Version 1.0, Volume 1 (ISBN 0-201-56788-1), Version 1.0, Volume 2 (ISBN 0-201-60845-6), and "Unicode Technical Report #4, The Unicode Standard, Version 1.1" (available from The Unicode Consortium, and soon to be published by Addison-Wesley).

[ISO 10646] ISO/IEC 10646-1:1993(E) Information Technology--Universal Multiple-octet Coded Character Set (UCS).

[UTF-7] Goldsmith, D., and M. Davis, "UTF-7: A Mail Safe Transformation Format of Unicode", RFC 1642, Taligent, Inc., July 1994.

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

[ISO-8859] Information Processing -- 8-bit Single-Byte Coded Graphic Character Sets -- Part 1: Latin Alphabet No. 1, ISO 8859-1:1987. Part 2: Latin alphabet No. 2, ISO 8859-2, 1987. Part 3: Latin alphabet No. 3, ISO 8859-3, 1988. Part 4: Latin alphabet No. 4, ISO 8859-4, 1988. Part 5: Latin/Cyrillic alphabet, ISO

8859-5, 1988. Part 6: Latin/Arabic alphabet, ISO 8859-6, 1987. Part 7: Latin/Greek alphabet, ISO 8859-7, 1987. Part 8: Latin/Hebrew alphabet, ISO 8859-8, 1988. Part 9: Latin alphabet No. 5, ISO 8859-9, 1990.

[RFC822] Crocker, D., "Standard for the Format of ARPA Internet Text Messages", STD 11, RFC 822, UDEL, August 1982.

[RFC-1521] Borenstein N., and N. Freed, "MIME (Multipurpose Internet Mail Extensions) Part One: Mechanisms for Specifying and Describing the Format of Internet Message Bodies", RFC 1521, Bellcore, Innosoft, September 1993.

[RFC-1522] Moore, K., "Representation of Non-Ascii Text in Internet Message Headers" RFC 1522, University of Tennessee, September 1993.

[UTF-8] X/Open Company Ltd., “File System Safe UCS Transformation Format (FSS_UTF)”, X/Open Preliminary Specification, Document Number: P316. This information also appears in Unicode Technical Report #4, and in a forthcoming annex to ISO/IEC 10646.

Authors’ Addresses

David Goldsmith
Taligent, Inc.
10201 N. DeAnza Blvd.
Cupertino, CA 95014-2233

Phone: 408-777-5225
Fax: 408-777-5081
EMail: david_goldsmith@taligent.com

Mark Davis
Taligent, Inc.
10201 N. DeAnza Blvd.
Cupertino, CA 95014-2233

Phone: 408-777-5116
Fax: 408-777-5081
EMail: mark_davis@taligent.com