Internet Draft
<draft-hoffman-widetext-01.txt>
December 12, 1998

Paul Hoffman
Internet Mail Consortium

Registration for the "widetext" Media Type

Status of this Memo

1. Introduction

This document defines a new MIME top-level media type, "widetext", which can be used to carry text that employs the UTF-16 character encoding scheme. The use of the "widetext" media type is limited to text-like MIME bodies that cannot be represented using the "text" media type.

1.1 Terminology

This document uses the same definitions for "type" and "top-level" that are used in the MIME media types document [MIMETYPES].

The internationalization community has a variety of definitions for many terms that have to do with characters. The following definitions are used in this document:

- A "character set" (more precisely called a "coded character set" or "CCS") is a mapping from a set of abstract characters to a set of integers. Examples of coded character sets include ISO 10646, US-ASCII, and the ISO 8859 series.

- A "character encoding scheme" or "CES" is a mapping from one or more coded character sets to a set of octets. Some CESs are associated with a single CCS; for example, UTF-16 applies only to ISO 10646. Other CESs, such as ISO 2022, are associated with many CCSs.

- A "charset" is a method of mapping a sequence of octets to a sequence of abstract characters. One way to construct a charset is to combine a CES with one or more CCSs.

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 RFC 2119 [MUSTSHOULD].

2. Need for the "widetext" type

[MIMETYPES] describes the purpose for the "text" type. Section 4.1 of that specification says:

> The "text" media type is intended for sending material which is principally textual in form.

However, not all character encoding schemes can be represented in "text" body parts. Section 4.1.1 of that specifications says:

> The canonical form of any MIME "text" subtype MUST always represent a line break as a CRLF sequence. Similarly, any occurrence of CRLF in MIME "text" MUST represent a line break. Use of CR and LF outside of line break sequences is also forbidden.

This means that a CES used with the "text" type must assure that the octets with the values 0x0D (CR) and 0x0A (LF) must never appear by themselves, and when they appear in the sequence 0x0D0A they must indicate an line break. Some popular CESs do not conform to this requirement.

In particular, the UTF-16 CES has many characters with bare 0x0D and 0x0A octets. The UTF-16 CES is optionally used by some document formats such as XML [XML].

Note that the "widetext" media type is being defined for the first time in this specification, whereas the "text" media type has been defined for many years and is deployed in every MIME agent. It is much more likely that the receiver of a MIME message will have an agent that understands the "text" type than one that the "widetext" type.

Thus, if the creator of a MIME body part has a choice, he or she should preferentially create a "text" type instead of a "widetext" type, even if they have to change from one CES to another (as long as that is allowed by the format requirements of the object). The only time a creator should use the "widetext" type is when they cannot use a "text" type due to the need to use a CES that cannot be used with the "text" type.

3. Definition of the "widetext" type

The "widetext" media type MUST only be used for sending material which is

principally textual in form and uses the UTF-16 CES, as defined in [ISO-10646]. (Note that other CESs that can be used with the "widetext" media type may be specified in the future.) A "charset" parameter MAY be used to indicate the character set of the body text for "widetext" subtypes.

It is noteworthy that the same set of characters is defined by the Unicode standard [UNICODE], which further defines additional character properties and other application details of great interest to implementors. Up to the present time, changes in Unicode and amendments to ISO/IEC 10646 have tracked each other, so that the character repertoires and code point assignments have remained in sync. The relevant standardization committees have committed to maintain this very useful synchronism.

3.1 Representation of line breaks

The definition of the line break characters in the canonical form of any subtype of "widetext" is explicitly undefined in this specification. Any charset that is used with a "widetext" subtype MUST have a method for indicating the ends of text lines.

3.2 Charset parameter

The "charset" parameter for the "widetext" type is similar to that for the "text" type. There are two significant differences:

- There is no default character set for the "widetext" type. This is a significant difference from the "text" type. In the "text" type, there are enough restrictions that you can still perform many operations even if you don't recognize the charset. This is not true of text in the "widetext" type. Therefore, "widetext" body that has no "charset" parameter SHOULD be treated as application/octet-stream.

- At the time of this writing, the only valid values for the "charset" parameter are "UTF-16", "UTF-16BE", and "UTF-16LE", as defined in [UTF16]. Note that each of these charsets have their byte order specified in the charset definition. Other valid values for the "charset" parameter may be registered in the future.

3.3 Default display semantics

For unrecognized subtypes in a known character set, a MIME displaying program MAY offer to display the text uninterpreted and MUST have the ability to save the text to a file (after removing any transfer encodings).

3.4 Encoding issues

UTF-16 text requires a binary-safe transport. Before sending a widetext object over a 7-bit or 8-bit transport, the sender SHOULD use Base64 transfer encoding.

3.5 Media requirements

The "widetext" type is used when the recipient is expected to have a processor to interpret UTF-16, and additionally have a display or printer with facilities that render ISO 10464.

4. Subtypes of "widetext"

The "text" type has many subtypes that have been defined. Some of the subtypes for "text" also apply for "widetext", while others do not. Registrations for all subtypes appear in Appendix A.

Note that the "widetext" type does not inherit any subtypes from the "text" type. All definitions of "widetext" subtypes must be specific to the "widetext" type.

Unrecognized subtypes of "widetext" should be treated as subtype "plain" as long as the MIME implementation knows how to handle the charset. Unrecognized subtypes which also specify an unrecognized charset should be treated as "application/octet-stream".

It is permitted to have a subtype of "widetext" that is not present in "text" and vice versa. If a subtype name is registered under both "widetext" and "text", the semantics MUST NOT differ in any way other than in the charsets that are permitted. If a subtype is available in both "widetext" and "text", an agent which generates the "widetext" form SHOULD be capable of generating the "text" form with the UTF-8 charset.

The canonical form for each subtype of "widetext" is lines ending with the character sequence "CARRIAGE RETURN" "LINE FEED" (0x000D 0x000A). Bare "CARRIAGE RETURN" (0x000D) or "LINE FEED" (0x000A) characters SHOULD NOT appear in any subtype of "widetext".

4.1 widetext/plain

The simplest and most important subtype of "widetext" is "plain". This indicates plain text that does not contain any formatting commands or directives. Plain text is intended to be displayed "as-is", that is, no interpretation of embedded formatting commands, font attribute specifications, processing instructions, interpretation directives, or content markup should be necessary for proper display.

In "widetext/plain", the character sequence "CARRIAGE RETURN" "LINE FEED" (0x000D 0x000A) is equivalent to the character "LINE SEPARATOR" (0x2028). A program creating "widetext/plain" text from primitive characters SHOULD use "CARRIAGE RETURN" "LINE FEED" instead of "LINE SEPARATOR". "widetext/plain" is permitted to carry columnar data such as formatted plain text tables intended for a fixed-width font display.

4.2 widetext/paragraph

The "paragraph" subtype of "widetext" is similar to the "plain" subtype,

except that it can be used for text that is in paragraph form using ISO 10646 **paragraph marks. Specifically:**

- the character sequence "CARRIAGE RETURN" "LINE FEED" (0x000D 0x000A) is equivalent to the character "PARAGRAPH SEPARATOR" (0x2029) - no column alignment or fixed-width display is presumed

4.3 Other allowed and disallowed subtypes

At the time of this specification, the following subtypes have been registered for the "text" type (this list excludes subtypes in the "prs." and "vnd." namespaces). Each subtype of "text" is analyzed for its ability to be used as a subclass of "widetext".

4.3.1 Additional subtypes

The "widetext" subtypes "html", "sgml", and "xml" are defined in Appendix A**. Other registrations for subtypes to "widetext" may appear in the future,** as long as they conform to the requirements in this specification.

4.3.2 Disallowed subtypes

directory -- MUST NOT be used as a subclass of widetext. Section 5.8.1 of RFC 2425 requires CRLFs for line terminators.

enriched -- MUST NOT be used as a subclass of widetext. RFC 1896 specifies that multi-byte character sets have to address internal conversion to an ASCII-compatible character set for markup.

rfc822-headers -- MUST NOT be used as a subclass of widetext. Only used to encode RFC 822 headers, which always use US-ASCII.

richtext -- MUST NOT be used as a subclass of widetext. This subtype is little used and may be obsolete.

rtf -- MUST NOT be used as a subclass of widetext. RTF uses only 7 bits per octet.

tab-separated-values -- MUST NOT be used as a subclass of widetext. This subtype is little used and may be obsolete.

uri-list -- MUST NOT be used as a subclass of widetext. URIs are only defined in US-ASCII.

5. Security considerations

The introduction of the "widetext" media type does not introduce any inherent security issues. However, using the UTF-16 charset definitely does introduce security issues, and those issues are covered in [UTF16].

6. References

[ISO-10646] ISO/IEC 10646-1:1993. International Standard -- Information technology -- Universal Multiple-Octet Coded Character Set (UCS) -- Part 1: Architecture and Basic Multilingual Plane. Twelve amendments and two technical corrigenda have been published up to now. UTF-16 is described in Annex Q, published as Amendment 1. Many other amendments are currently at various stages of standardization.

[MIMETYPES] N. Freed, N. Borenstein, "MIME Part Two: Media Types", RFC 2046, November 1996.

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

[UTF16] "UTF-16, an encoding of ISO 10646", draft in progress, draft-hoffman-utf16-xx.txt.

[UNICODE] The Unicode Consortium, "The Unicode Standard -- Version 2.1", Unicode Technical Report #8.

[XML] T. Bray, J. Paoli, C. M. Sperberg-McQueen, "Extensible Markup Language (XML)", World Wide Web Consortium Recommendation REC-xml-19980210, <http://www.w3.org/TR/1998/REC-xml-19980210>.

7. Acknowledgments

Chris Newman contributed a great deal of editing and writing to the early drafts of this document. Other significant contributors include:
Keith Moore
Martin Duerst
Ned Freed

8. Author's address

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

9. Changes from -00 to -01

Small editorial changes throughout.

3.2: Added a bunch of text to the first bullet to explain why you should default to application/octet-stream if there is not charset given.

4: Added the second paragraph. In the (now) fourth paragraph, downgraded

the MUST to SHOULD.

4.3.2: Removed "css" from the beginning of the list. Also updated the reasoning for "rtf" from the current MIME registration.

A: Set all the Macintosh type codes to "none".

A.6: Added this because CSS does all UTF-16.

A. Subtype registrations

A.1 widetext/plain

To: ietf-types@iana.org
Subject: Registration of MIME media type widetext/plain

MIME media type name: widetext

MIME subtype name: plain

Required parameters: none

Optional parameters: charset

Encoding considerations: All allowed charsets require transfer encoding for 7-bit or 8-bit environments.

Security considerations: See security section of this specification.

Interoperability considerations: Text in "widetext/plain" can be converted to "text/plain" only for applications that allow UTF-8, and only if the input text uses the same line-ending semantics as "text/plain".

Published specification: This specification

Applications which use this media type: Any application that requires the use of UTF-16.

Additional information:

  Magic number(s): none
  File extension(s): .txt
  Macintosh File Type Code(s): none

Person & email address to contact for further information:
Paul Hoffman <phoffman@imc.org>

Intended usage: COMMON

Author/Change controller:
Paul Hoffman <phoffman@imc.org>

Other requirements for "widetext/plain" are given in the main body of this

specification.

A.2 widetext/paragraph

To: ietf-types@iana.org
Subject: Registration of MIME media type widetext/paragraph

MIME media type name: widetext

MIME subtype name: paragraph

Required parameters: none

Optional parameters: charset

Encoding considerations: All allowed charsets require transfer encoding for 7-bit or 8-bit environments.

Security considerations: See security section of this specification.

Interoperability considerations: Text in "widetext/paragraph" can be converted to "text/plain" only for applications that allow UTF-8, and only if the input text uses the same line-ending semantics as "text/plain".

Published specification: This specification

Applications which use this media type: Any application that requires the use of UTF-16.

Additional information:

  Magic number(s): none
  File extension(s): .txt
  Macintosh File Type Code(s): none

Person & email address to contact for further information:
Paul Hoffman <phoffman@imc.org>

Intended usage: COMMON

Author/Change controller:
Paul Hoffman <phoffman@imc.org>

Other requirements for "widetext/paragraph" are given in the main body of this specification.

A.3 widetext/html

To: ietf-types@iana.org
Subject: Registration of MIME media type widetext/html

MIME media type name: widetext

MIME subtype name: html

Required parameters: none

Optional parameters: charset

Encoding considerations: All allowed charsets require transfer encoding for 7-bit or 8-bit environments.

Security considerations: See security section of this specification.

Interoperability considerations: Text in "widetext/html" can be converted to "text/html".

Published specification: HTML is defined in RFC 1866. More recently, HTML has been defined by the W3C at <http://www.w3.org/TR/REC-html40>.

Applications which use this media type: HTML applications that require the use of UTF-16.

Additional information:

  Magic number(s): none
  File extension(s): .htm or .html
  Macintosh File Type Code(s): none

Person & email address to contact for further information:
Paul Hoffman <phoffman@imc.org>

Intended usage: COMMON

Author/Change controller:
Paul Hoffman <phoffman@imc.org>

A.4 widetext/sgml

To: ietf-types@iana.org
Subject: Registration of MIME media type widetext/sgml

MIME media type name: widetext

MIME subtype name: sgml

Required parameters: none

Optional parameters: charset, SGML-bctf, SGML-boot

Encoding considerations: All allowed charsets require transfer encoding for 7-bit or 8-bit environments.

Security considerations: See security section of this specification.

Interoperability considerations: Text in "widetext/sgml" can be converted

to "text/sgml" and "application/sgml".

Published specification: This registration is based on RFC 1874.

Applications which use this media type: SGML applications that require the use of UTF-16.

Additional information:

  Magic number(s): none
  File extension(s): none
  Macintosh File Type Code(s): none

Person & email address to contact for further information:
Paul Hoffman <phoffman@imc.org>

Intended usage: COMMON

Author/Change controller:
Paul Hoffman <phoffman@imc.org>

A.5 widetext/xml

To: ietf-types@iana.org
Subject: Registration of MIME media type widetext/xml

MIME media type name: widetext

MIME subtype name: xml

Required parameters: none

Optional parameters: charset

Encoding considerations: All allowed charsets require transfer encoding for 7-bit or 8-bit environments.

Security considerations: See security section of this specification.

Interoperability considerations: Text in "widetext/xml" can be converted to "text/xml" and "application/sgml".

Published specification: This registration is based on RFC 2376.

Applications which use this media type: XML applications that require the use of UTF-16.

Additional information:

  Magic number(s): none
  File extension(s): .xml
  Macintosh File Type Code(s): none

Person & email address to contact for further information:
Paul Hoffman <phoffman@imc.org>

Intended usage: COMMON

Author/Change controller:
Paul Hoffman <phoffman@imc.org>

A.6 widetext/css

To: ietf-types@iana.org
Subject: Registration of MIME media type widetext/css

MIME media type name: widetext

MIME subtype name: css

Required parameters: none

Optional parameters: charset

Encoding considerations: All allowed charsets require transfer encoding for 7-bit or 8-bit environments.

Security considerations: See security section of this specification.

Interoperability considerations: Text in "widetext/css" can be converted to "text/css".

Published specification: This registration is based on RFC 2318.

Applications which use this media type: CSS applications that require the use of UTF-16.

Additional information:

  Magic number(s): none
  File extension(s): .css
  Macintosh File Type Code(s): none

Person & email address to contact for further information:
Paul Hoffman <phoffman@imc.org>

Intended usage: COMMON

Author/Change controller:
Paul Hoffman <phoffman@imc.org>