Network Working Group
Internet Draft
draft-palme-text-html-01.txt
Category-to-be: Standard
Expires July 1996

Jacob Palme
Stockholm University/KTH
Sweden

January 1996

The Text/HTML content type and the Content-Location MIME header

or

Sending HTML documents via MIME e-mail

Status of this Memo

This document is an Internet-Draft. 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.''

To learn the current status of any Internet-Draft, please check the ``1id-abstracts.txt'' listing contained in the Internet-Drafts Shadow Directories on ftp.is.co.za (Africa), nic.nordu.net (Europe), munnari.oz.au (Pacific Rim), ds.internic.net (US East Coast), or ftp.isi.edu (US West Coast).

This memo provides information for the Internet community. This' memo does not specify an Internet standard of any kind, since this document is mainly a compilation of information taken from other RFC-s. Distribution of this memo is unlimited.

Abstract

This memo specifies how to send HTML-formatted documents in Internet mail. The memo particularly addresses the issue of handling of hyperlinks in HTML documents referring to other body parts in the same message. In order to do this, the memo introduces one new MIME content-header with the name "Content-Location".

Differences from Previous Version

The postscript (.ps) version of this draft shows the differences between version 00 and 01 through underscoring and strikethru markup.

This document has been revised based on the discussions in the ietf-types and mhtml mailing lists and in the BOF at the Dallas IETF meeting in December 1995.

Use of the Content-Base header has been introduced. The "linking" parameter has been removed and replaced with use of the Content-Base parameter.

Use of the Content-Disposition header has been replaced with use of the "Content-Base: FILE" och "Content-Location" ~~and one new attribute to the MIME header "Content-Type: Text/HTML" with the name "linking".~~ headers.

Information on the new mailing list for further discussions of this ietf draft has been added.

Syntax for embedding URI-s in MIME headers has been added, copied from [URLBODY].

Security considerations for implementations using proxy servers has been added.

A temporary annex on implementation has been added. This annex might be removed in the final version of this standard.

Table of Contents

Further Discussion

Further discussion on this memo should be sent to the mailing list mhtml@segate.sunet.se.

To subscribe to this list, send a message to
listserv@segate.sunet.se
which contains the text
sub mhtml <your name (not your e-mail address)>

Archives of this list are available by anonymous ftp from
ftp://segate.sunet.se
The archives are also available by e-mail. Send a message to listserv@segate.sunet.se with the text "index mhtml" to get a list of the archive files, and then a new message "get <file name>" to retrieve the archive files.
or
get mhtml digest

1. Introduction

The HTML format is a very common format for documents in the Internet, and there is an obvious need to be able to send documents in this format in e-mail [SMTP, RFC822]. The "text/html; version=2.0" media type is defined in [HTML2]. This memo gives additional specifications and advice on how to use the text/html media type as a Content-Type in MIME ~~[MIME]~~ [MIME1] e-mail messages.

2. Terminology

Most of the terms used in this memo are defined in other RFC-s. For example, URL is defined in [URL], URI, absolute URI, and relative URI is defined in [HTML2].

3. The Content-Location MIME Content Header

An additional MIME heading field is defined with the name "Content-Location". This header field can occur in any MIME message heading or content heading. Its value ~~is~~ can be an absolute or relative URI.

A relative URI in the Content-Location header is only allowed if there is also a Content-Base header (as defined in [RELURL]) specifying the base for the relative URI.

~~URI. The data below the current heading, and the data~~ This header is used to indicate that the data sent under this heading is also retrievable, in identical formal, through normal use of this URI. Thus, the information sent in the message can be seen as a cached version of the original data. This header is only permitted if the data is actually retrievable through use of this ~~URI, should be identical.~~ URI.

In practice, at present only those URI-s which are URL-s are used, but it is anticipated that other forms of URI-s will in the future be used. This heading is similar to the Location header as defined in [HTTP].

The syntax for the new heading field is, using the syntax definition tools from [RFC822]:

    content-location ::= "Content-Location:" ~~<"> uri <">~~ URI-parameter

where ~~uri~~ URI is at present (November 1995) restricted to the syntax for URL-s as defined in [URL]. This syntax will be widened when the definition of the URI syntax becomes more stable. The URI must encoded in a format which allows for splitting of long URI-s into more than one line. This is done using the following syntax, copied from [URLBODY]:

```
URL-parameter := <"> URL-word *(*LWSP-char URL-word) <">

URL-word := token
              ; Must not exceed 40 characters in length
```

The syntax of an actual URL string is given in [URL]. URL strings can be of any length and can contain arbitrary character content. This presents problems when URLs are embedded in MIME body part headers that are wrapped according to RFC 822 rules. For this reason they are transformed into a URL-parameter for inclusion in a message/external-body content-type specification as follows:

(1) A check is made to make sure that all occurrences of SPACE, CTLs, double quotes, backslashes, and 8-bit characters in the URL string are already encoded using the URL encoding scheme specified in RFC 1738. Any unencoded occurrences of these characters must be encoded. Note that the result of this operation is nothing more than a different representation of the original URL.

(2) The resulting URL string is broken up into substrings of 40 characters or less.

(3) Each substring is placed in a URL-parameter string as a URL-word, separated by one or more spaces. Note that the enclosing quotes are always required since all URLs contain one or more colons, and colons are tspecial characters [RFC 1521].

Extraction of the URL string from the URL-parameter is even simpler: The enclosing quotes and any linear whitespace are removed and the remaining material is the URL string.

Note: This header is similar to the Location header defined in [HTTP].

4. Parameters for the Content-Type: Text/HTML

The optional "version" parameter for the Content-Type: Text/HTML indicates the version of HTML used, with "2.0" as default value.

~~The new optional parameter "linking" defined in chapter 10 below.~~

5. Use of Relative URL-s in Text/HTML Contents

The use of relative URL-s in Content-Type: Text/HTML ~~SHOULD~~ should never be used except in one of the following three ~~cases:~~ cases (in order of priority, if more than one of them are present, the first-listed applies)

(a) There is a BASE element in the HTML document which resolves the relative URL into a non-relative URL.

~~(b) The relative URL refers to another body part in the same message as defined in chapter 9.2 Filename-method below.~~

~~(c) In special cases, where the sender and recipient have prior agreement on the resolution of relative URL-s.~~

(b) There is a Content-Base header (as defined in [RELURL]), giving the base to be used.

(c) There is a Content-Location of the Text/HTML which can then serve as the base.

6. Use of the Content-Type: Multipart/related

A message can contain one or more Text/HTML body parts and also contain as separate body parts, data, to which hyperlinks (as defined in [HTML2]) in the Text/HTML body part refers.

~~If this is done, it is recommended to use the multipart/related Content-Type~~ Such embedded linked parts must, together with the Text/HTML body part, be enclosed within a Multipart/Related body part as defined in [REL].

The root (as defined in [REL]) should then be of the Content-Type: Text/HTML.

Such an embedded linked part can itself be a Multipart/related body parts including its own linked objects.

7. Use of the Content-type: Multipart/alternative

If the message is sent to recipients, all of which may not have mailers capable of handling the Text/HTML content-type, then the Content-Type: Multipart/Alternative ~~[MIME]~~ [MIME1] can be used, for example with Content-Type: Text/plain as the first choice, and Content-Type: Text/HTML as the second choice.

8. Combination of the Content-Types: Multipart/related and Multipart/alternative.

Both the Content-type: Multipart/related, as defined in chapter 6 above and the Content-Type: Multipart/alternative, as defined in chapter 7 above can be combined in the same message. It is then recommended to put the Multipart/alternative inside the Multipart/related. Note that ~~Example:~~

if this is done, a start parameter to the Content-Type: ~~Multipart/related; boundary="boundary-example-1"~~ Multipart/ ~~type=Text/HTML~~
related is necessary, as shown by the example below.

Example:

```
   Content-Type: Multipart/related; boundary="boundary-example-1";
                 type=Text/HTML; start=content-id-example@example.host

      --boundary-example 1
      Content-Type: MULTIPART/ALTERNATIVE
      Boundary: boundary-example-2

         --boundary-example-2
         Content-Type: Text/plain

         ... plain text version of the document for recipients
         whose mailers cannot handle Text/HTML ...

         --boundary-example-2
         Content-Type: Text/HTML
         Content-ID: content-id-example@example.host

         ... text of the HTML document ...

         --boundary-example-2--
      --boundary-example-1
      Content-Type: Image/GIF

      ... a body part, to which the HTML document has a link  ...
      --boundary-example-1--
```

9. Format of Links to Other Body Parts

A Text/HTML body part may contain hyperlinks to documents which are included as other body parts in the same ~~message.~~ message and within the same multipart/related content. Three ways to do this is specified in this memo:

~~9.1 Location-method: Use of the~~ 9.1 General Location-Method: Identical URI-s in Content-Location ~~field~~ headers

With this method, All URI-s in the Text/HTML document SHOULD be ~~non-relative~~ absolute URI-s as defined in ~~[HTML2], and it~~ [HTML2] or relative URI-s relative to a surrounding Content-Base header. It SHOULD be possible to use these URI-s to retrieve the referred document using the protocol defined for retrieval of this particular URL scheme in [URL] (subject to access control).

For each distinct URI in the Text/HTML document, which refers to data which is sent in the same MIME message, there SHOULD be a ~~separata~~ separate body part ~~in~~ within the multipart/related part of the message containing this data. Each such body part SHOULD contain a Content-Location heading field, and the ~~value of~~ string in this field SHOULD be identical to the URI as used in the Text/HTML document.

Note: By identical string is not meant equivalent URI-s but actually identical URI strings.

The receiving mailer can then resolve the hyperlink either by using the URI in the normal way, or by using the data in the body part whose Content-Location contains the same URI.

Example with absolute URI-s:

```
   Content-Type: Multipart/related; boundary="boundary-example-1";
                 type=Text/HTML

      --boundary-example 1
      Content-Type: Text/HTML

      ... text of the HTML document, which might contain a hyperlink
      to the other body part, for example through a statement such as:


      <IMG SRC="http://www.dsv.su.se/images/DSV-logo.gif">

      --boundary-example-1
      Content-Type: Image/GIF
      Content-Location: "http://www.dsv.su.se/images/DSV-sign-eng.gif"
"http://www.dsv.su.se/images/logo.gif"

      --boundary-example-1--
```

9.2 Filename-method: Use of file names

Example with relative URI-s:

```
   Content-Base: http://www.dsv.su.se
   Content-Type: Multipart/related; boundary="boundary-example-1";
                 type=Text/HTML

      --boundary-example 1
      Content-Type: Text/HTML

      ... text of the HTML document, which might contain a hyperlink
      to the other body part, for example through a statement such as:
      <IMG SRC="/images/logo.gif">

      --boundary-example-1
      Content-Type: Image/GIF
      Content-Location: "/images/logo.gif"

      --boundary-example-1--
```

9.2 Filename-Method: Use of virtual File Names

This method is a special case of the Location-Method described in section 9.1, but also differs in that it may be used even if the enclosed parts are not retrievable from other places than the body parts included in the message.

With this method, the hyperlink URIs to other body parts in the same message in the Text/HTML document SHOULD have a very simple format. This simple format is relative URL-s of the form

```
   relative-url ::= 1ALPHA 0#7ALPHADIGIT [ "." 1#3ALPHADIGIT ]

   ALPHADIGIT ::= ALPHA / DIGIT
```

i.e. 1-8 characters plus 0-3 extension characters, only using Ascii letters and digits and beginning with a letter.

The choice of this simple format is to match permitted file name formats in most operating systems in wide use today.

For each distinct URI in the Text/HTML document, which refers to data which is sent in the same MIME message, there should be a separate body ~~part~~ part, within the same multi-part/related content in the ~~message~~ message, containing this data. Each such body part SHOULD contain a ~~Content-Disposition header [RFC 1806] with a filename parameter. The value of this filename parameter~~ Content-Location header. The string in this Content-Location header should be identical to the relative URI as used in the Text/HTML document.

~~The value of the Content-Disposition header should be "inline" if the URI in the Text/HTML document is used for an inline HTML element, such~~

~~as an <IMG> element, and should be "attachment" if the URI in the Text/HTML document i used for a hyperlink to a document which is activated at the request of the recipient, such as an <A> element.~~

Note: This method does not require that the body parts are actually stored in files in the recipient computer. The receiving mailer may choose to implement this method by storing the individual body parts in files with the virtual file name, or may choose other implementation methods.

Example:

```
   Content-Base: "FILE:"
   Content-Type: Multipart/related; boundary="boundary-example-1";
                 type=Text/HTML

      --boundary-example 1
      Content-Type: Text/HTML

      ... text of the HTML document, which might contain a hyperlink
      to the other body part, for example through a statement such as:
      <IMG SRC="logo.gif">

      --boundary-example-1
      Content-Type: Image/GIF
      Content-Location: "logo.gif"

      --boundary-example-1--
```

9.3 CID-method: Use of CID URL-s

With this method, the hyperlink URIs to other body parts in the same message in the Text/HTML document SHOULD be CID (Content-ID) URL-s as defined in [URL] and [MIDCID].

For each distinct URI in the Text/HTML document, which refers to data which is sent in the same MIME message, there should be a separate body part in the message containing this data. Each such body part SHOULD have a Content-ID header ~~[MIME].~~ [MIME1]. The value of this Content-ID header should be identical to the CID as used in the Text/HTML document.

Example:

```
   Content-Type: Multipart/related; boundary="boundary-example-1";
                 type=Text/HTML

      --boundary-example 1
      Content-Type: Text/HTML

      ... text of the HTML document, which might contain a hyperlink
      to the other body part, for example through a statement such as:
      <IMG SRC="cid:sign-eng*jpalme@dsv.su.se">
```

```
--boundary-example-1
Content-Type: Image/GIF
Content-ID: sign-eng*jpalme@dsv.su.se

--boundary-example-1--
```

Note: Content-ID-s should be globally unique. It is not permitted to make them unique only within this message or within this multipart/related.

9.4 Recommended Choice of Method:

A Text/HTML content may always, in addition to the use the methods described in this chapter of this memo, contain URI-s only resolvable using the method defined for this particular URI scheme, and not referring to any data in separate body parts of the same message.

Method	Body part identification method	Recommendation
Virtual file name method	File name in Content-Location header	Recommended as the primary choice, to be used whenever possible. ~~For private documents or documents not retrievable.~~
General Content-Location method	Content-Location header	~~When the referred document is~~ Recommended if existing HTML documents are to be sent unchanged, but only if the referred-to document(s) are publicly available and retrievable using the scheme used in the URI.
CID method	Content-ID header	For experimental use between consenting partners.

10. Indication of Method Used

~~There should be an additional optional parameter to the Content-Type: Text/HTML header, with the name "linking" and the syntax:~~

~~linking ::= "linking=" linkmethod~~

~~where linkmethod can have the following values:~~

~~external~~	~~Only use of absolute URI-s and retrieval using the scheme defined for this URI, and not containing any URI referring to other body parts in the same essage.~~
~~location~~	~~Location-method as defined in chapter 9.1.~~
~~filename~~	~~Filename-method as defined in chapter 9.2.~~
~~cid~~	~~CID-method as defined in chapter 9.3.~~

~~Default value if this parameter is omitted is "external".~~

Which of the methods above used is indicated by the value of the surrounding Content-Base header:

Method	Indicated by:
Virtual file name method	Content-Base: FILE: as defined in [URL]
CID method	Content-Base: CID as defined in [CID]
General Content-Location method	Any other Content-Base or no Content-Base specified

(??) Should LOCAL-FILE, as defined in [MIME2] be used instead of FILE as defined in ? Or should something new, such as "LOCAL" or "VIRTUAL FILE" be used to clarify that no real file storage is necessary?

11. Content-Disposition header

Information in the Content-Disposition header (as defined in [CONDISP]) on individual body parts within a multipart/related is ignored. Receiving mailers which are not capable of handling the multipart/related header, and which thus by default handles this header as if it was multipart/mixed, can however make use of information in the Content-Disposition header.

12. Sending forms in e-mail

When an e-mail message contains an HTML form, then the default for ACTION (as defined in [HTML2] section 8.1.1) should be replying by e-mail to the From: or Reply-To address of the message containing the form, and not, as specified in [HTML2], the base URI of the document.

13. Encoding Considerations

There are two recommended ways to encode 8-bit characters in Text/HTML contents:

(1) Let the charset of the content part be iso-8859-1, and encode the content with the quoted-printable encoding method.

(2) Let the charset of the content part be us-ascii, and encode non-us-ascii characters in the text using the Data character encoding defined in [HTML2].

Both these encoding methods are permitted, and they can also be mixed in the same document. Recipients must be capable of handling both encoding alternatives. However, it is recommended that encoding method (2) above is used when sending Text/HTML messages.

If only method (2) is used, the charset parameter should be "us-ascii".

If method (1), or a mixture of method (1) and method (2) is used, the charset parameter should be "iso-8859-1".

14. Security Considerations

There is a potential security risk if the Content-Location: heads a body part whose data is not identical to that retrievable using the URI in the Content-Location. To reduce this risk, it might be unsuitable to cache the data in such a way that the cached data can be used for retrieval of this URL from other documents than those included in the same message as the Content-Location header.

One way of implementing messages with linked body parts is to handle the linked body parts in a combined mail and WWW proxy server. The mail client is only given the start body part, which it turns over to a web browser. This web browser requests the linked parts in the normal way, but these requests are intercepted by the proxy server. If this method is used, and if the combined server is used by more than one user, then methods must be employed to ensure that body parts of a message to one person is not retrievable by another person. Use of passwords (also known as tickets or magic cookies) is one way of achieving this.

15. Acknowledgements

Harald Tveit Alvestrand, Richard Baker, Al Gilman, Roy Fielding, Keith Moore, Ed Levinson, Al Gilman, Mark K. Joseph, Daniel LaLiberte, Valdis Kletnieks, Larry Masinter and several other people have helped me with preparing this memo. I alone take responsibility for any errors which may still be in the memo.

16. References

Temporary note: This list contains some references to Internet drafts. It is anticipated that these Internet drafts will become RFC-s before this memo. The references will then in this memo be changed to refer to the corresponding RFC instead.

Ref.	Author, title
~~[CIDMID]~~	~~E. Levinson: "Content-ID and Message-ID Uniform Resource Locators", <draft-levinson-cid-01.txt>, October 1995.~~
[CID]	E. Levinson: "Message/External-Body Content-ID Access Type", RFC 1873, December 1995.
[CONDISP]	**R. Troost, S. Dorner: "Communicating Presentation Information in Internet Messages: The Content-Disposition Header", RFC 1806, June 1995.**
[HOSTS]	R. Braden (editor): "Requirements for Internet Hosts -- Application and Support", STD-3, RFC 1123, October 1989.
[HTTP]	T. Berners-Lee, R. Fielding, H. Frystyk: "Hypertext Transfer Protocol - - HTTP/1.0", <draft-ietf-http-v10-spec-04.txt>, April 1996.
[MIME1]	N. Borenstein & N. Freed: "MIME (Multipurpose Internet Mail Extensions) Part One: Mechanisms for Specifying and Describing the Format of Internet Message Bodies", RFC 1521, Sept 1993.
[MIME2]	N. Borenstein & N. Freed: "Multipurpose Internet Mail Extensions (MIME) Part Two: Media Types". draft-ietf-822ext-mime-imt-02.txt, December 1995.

[NEWS] M.R. Horton, R. Adams: "Standard for interchange of USENET messages", RFC 1036, December 1987.

[REL] Harald Tveit Alvestrand, Edward Levinson: "The MIME Multipart/Related Content-type", <draft-levinson-multipart-related-00.txt>, January 1995.

[RELURL] R. Fielding: "Relative Uniform Resource Locators", RFC 1808, June 1995.

[RFC822] D. Crocker: "Standard for the format of ARPA Internet text messages." STD 11, RFC 822, August 1982.

[SMTP] J. Postel: "Simple Mail Transfer Protocol", STD 10, RFC 821, August 1982.

[URL] T. Berners-Lee, L. Masinter, M. McCahill: "Uniform Resource Locators (URL)", RFC 1738, December 1994.

[URLBODY] N. Freed and Keith Moore: "Definition of the URL MIME External-Body Access-Type", draft-ietf-mailext-acc-url-01.txt, November 1995.

[HTML2] T. Berners-Lee, D. Connolly: "Hypertext Markup Language - 2.0", RFC 1866, November 1995.

17. Author's Address

Jacob Palme
Stockholm University and KTH
Electrum 230
S-164 40 Kista, Sweden

Phone: +46-8-16 16 67
Fax: +46-8-783 08 29
E-mail: jpalme@dsv.su.se

Annex A: Implementation methods

This annex is not part of the standards and is only included for informational purposes. This annex might be removed before making this memo into an IETF standard.

This standard has been intentionally written to be implementable both in cases where the web browser and e-mail program is combined, and when they are separate programs. Implementation is of course no problem if the web browser is combined with the e-mail client.

```
    +----------+                                      +---------+
    | Web      |                                      | Mail    |
    | browser  |                                      | client  |
    +-------+--+                                      +-+-------+
            |                                           |
         +--+-------------------------------+--+
         | +----------+  +--+  +--+                    |
         | | Start    |  |  |  |  | Related            |
         | | HTML     |  |  |  |  | body part          |
         | | document |  |  |  |  | parts              |
         | +----------+  +--+  +--+                    |
         +---------------------------------------+
```

If the web browser is separate from the e-mail client, the e-mail client might turn over the HTML body part to the web browser and ask it to display it. One way of doing this is to store the HTML body part in a file, and ask the web browser to display this file. If multipart/related is used, this can be implemented by storing all the body parts within the multipart/related in an otherwise empty folder/directory. With the virtual file name method described in section 9.2 above, this does not require any rewriting of the HTML text and is thus easy to implement, that is why the virtual file name is recommended as the primary method above.

```
    +----------+                                     +---------+
    | Web      |                                     | Mail    |
    | browser  |                                     | client  |
    +-------+--+                                     +-+-------+
            |                                          |
         +--+------------------------------+-+
         | +--------+  +--+  +--+                     |
         | | Trans- |  |  |  |  | Related             |
         | | lation |  |  |  |  | body part           |
         | | table  |  |  |  |  | parts               |
         | +--------+  +--+  +--+                     |
         +-------------------------------------+
```

With the general Content-Location methods, the web browser must in some way be instructed to retrieve the body parts from the received message. This can be done by a translation table, if the web browser has an API which allows for such a table.

```
    +---------+       +------------+       +---------+
    | Proxy   |       | Data base  |       | Mail    |
    | web     |-------| of cached  |-------| server  |
    | server  |       | objects    |       |         |
    +----+----+       +------------+       +----+----+
         |                                      |
    +----+----+                            +----+----+
    | Web      |                           | Mail    |
    | browser  |                           | client  |
    +--------+-+                           +-+-------+
             |                               |
         +---+-------------------------------+-+
         |          Start HTML object          |
         +-------------------------------------+
```

Other methods are to rewrite the HTML text before turning it over to the web browser, and to use a proxy web server, to which the web browser requests are sent, and which will then use the cached body parts instead of normal web retrieval from the network.

This comparison of two versions of this IETF draft was produced using the Docucomp II comparison program.