Network Working Group
Request for Comments: 2854
Obsoletes: 2070, 1980, 1942, 1867, 1866
Category: Informational

D. Connolly
World Wide Web Consortium (W3C)
L. Masinter
AT&T

June 2000

The 'text/html' Media Type

Status of this Memo

This memo provides information for the Internet community. It does not specify an Internet standard of any kind. Distribution of this memo is unlimited.

Copyright Notice

Abstract

This document summarizes the history of HTML development, and defines the "text/html" MIME type by pointing to the relevant W3C recommendations; it is intended to obsolete the previous IETF documents defining HTML, including RFC 1866, RFC 1867, RFC 1980, RFC 1942 and RFC 2070, and to remove HTML from IETF Standards Track.

This document was prepared at the request of the W3C HTML working group. Please send comments to www-html@w3.org, a public mailing list with archive at <http://lists.w3.org/Archives/Public/www-html/>.

1. Introduction and background

HTML has been in use in the World Wide Web information infrastructure since 1990, and specified in various informal documents. The text/html media type was first officially defined by the IETF HTML working group in 1995 in [HTML20]. Extensions to HTML were proposed in [HTML30], [UPLOAD], [TABLES], [CLIMAPS], and [I18N].

The IETF HTML working group closed Sep 1996, and work on defining HTML moved to the World Wide Web Consortium (W3C). The proposed extensions were incorporated to some extent in [HTML32], and to a larger extent in [HTML40]. The definition of multipart/form-data from [UPLOAD] was described in [FORMDATA]. In addition, a reformulation of HTML 4.0 in XML 1.0[XHTML1] was developed.

[HTML32] notes "This specification defines HTML version 3.2. HTML 3.2 aims to capture recommended practice as of early '96 and as such to be used as a replacement for HTML 2.0 (RFC 1866)." Subsequent specifications for HTML describe the differences in each version.

In addition to the development of standards, a wide variety of additional extensions, restrictions, and modifications to HTML were popularized by NCSA's Mosaic system and subsequently by the competitive implementations of Netscape Navigator and Microsoft Internet Explorer; these extensions are documented in numerous books and online guides.

2. Registration of MIME media type text/html

MIME media type name: text
MIME subtype name: html
Required parameters: none
Optional parameters:

charset
The optional parameter "charset" refers to the character encoding used to represent the HTML document as a sequence of bytes. Any registered IANA charset may be used, but UTF-8 is preferred. Although this parameter is optional, it is strongly recommended that it always be present. See Section 6 below for a discussion of charset default rules.

Note that [HTML20] included an optional "level" parameter; in practice, this parameter was never used and has been removed from this specification. [HTML30] also suggested a "version" parameter; in practice, this parameter also was never used and has been removed from this specification.

Encoding considerations:
See Section 4 of this document.

Security considerations:
See Section 7 of this document.

Interoperability considerations:
HTML is designed to be interoperable across the widest possible range of platforms and devices of varying capabilities. However, there are contexts (platforms of limited display capability, for example) where not all of the capabilities of the full HTML definition are feasible. There is ongoing work to develop both a modularization of HTML and a set of profiling capabilities to identify and negotiate restricted (and extended) capabilities.

Due to the long and distributed development of HTML, current practice on the Internet includes a wide variety of HTML variants. Implementors of text/html interpreters must be prepared to be "bug-compatible" with popular browsers in order to work with many HTML documents available the Internet.

Typically, different versions are distinguishable by the DOCTYPE declaration contained within them, although the DOCTYPE declaration itself is sometimes omitted or incorrect.

Published specification:
The text/html media type is now defined by W3C Recommendations; the latest published version is [HTML401]. In addition, [XHTML1] defines a profile of use of XHTML which is compatible with HTML 4.01 and which may also be labeled as text/html.

Applications which use this media type:
The first and most common application of HTML is the World Wide Web; commonly, HTML documents contain URI references [URI] to other documents and media to be retrieved using the HTTP protocol [HTTP]. Many gateway applications provide HTML-based interfaces to other underlying complex services. Numerous other applications now also use HTML as a convenient platform-independent multimedia document representation.

Additional information:

Magic number:
There is no single initial string that is always present for HTML files. However, Section 5 below gives some guidelines for recognizing HTML files.

File extension:
The file extensions 'html' or 'htm' are commonly used, but other extensions denoting file formats for preprocessing are also common.

Macintosh File Type code: TEXT

Person & email address to contact for further information:
Dan Connolly <connolly@w3.org>
Larry Masinter <lmm@acm.org>

Intended usage: COMMON

Author/Change controller:
The HTML specification is a work product of the World Wide Web Consortium's HTML Working Group. The W3C has change control over the HTML specification.

Further information:
HTML has a means of including, by reference via URI, additional resources (image, video clip, applet) within the base document. In order to transfer a complete HTML object and the included resources in a single MIME object, the mechanisms of [MHTML] may be used.

3. Fragment Identifiers

The URI specification [URI] notes that the semantics of a fragment identifier (part of a URI after a "#") is a property of the data resulting from a retrieval action, and that the format and interpretation of fragment identifiers is dependent on the media type of the retrieval result.

For documents labeled as text/html, the fragment identifier designates the correspondingly named element; any element may be named with the "id" attribute, and A, APPLET, FRAME, IFRAME, IMG and MAP elements may be named with a "name" attribute. This is described in detail in [HTML40] section 12.

4. Encoding considerations

Because of the availability within HTML itself for using character entity references, documents that use a wide repertoire of characters may still be represented using the US-ASCII charset and transported without encoding. However, transport of text/html using a charset other than US-ASCII may require base64 or quoted-printable encoding for 7-bit channels.

As with all MIME text subtypes, the canonical form of "text/html" must always represent a line break as a sequence of a CR byte value (0x0D) followed by an LF (0x0A) byte value. Similarly, any occurrence of such a CRLF sequence in "text/html" must represent a line break. Use of CR byte values and LF byte values outside of line break sequences is also forbidden. This rule applies regardless of the character encoding ('charset') involved.

Note, however, that the HTTP protocol allows the transport of data not in canonical form, and, in particular, with other end-of-line conventions; see [HTTP] section 3.7.1. This exception is commonly used for HTML.

HTML sent via email is still subject to the MIME restrictions; this is discussed fully in [MHTML] Section 10.

5. Recognizing HTML files

Almost all HTML files have the string "<html" or "<HTML" near the beginning of the file.

Documents conformant to HTML 2.0, HTML 3.2 and HTML 4.0 will start with a DOCTYPE declaration "<!DOCTYPE HTML" near the beginning, before the "<html". These dialects are case insensitive. Files may start with white space, comments (introduced by "<!--"), or processing instructions (introduced by "<?") prior to the DOCTYPE declaration.

XHTML documents (optionally) start with an XML declaration which begins with "<?xml" and are required to have a DOCTYPE declaration "<!DOCTYPE html".

6. Charset default rules

The use of an explicit charset parameter is strongly recommended. While [MIME] specifies "The default character set, which must be assumed in the absence of a charset parameter, is US-ASCII." [HTTP] Section 3.7.1, defines that "media subtypes of the 'text' type are defined to have a default charset value of 'ISO-8859-1'". Section 19.3 of [HTTP] gives additional guidelines. Using an explicit charset parameter will help avoid confusion.

Using an explicit charset parameter also takes into account that the overwhelming majority of deployed browsers are set to use something else than 'ISO-8859-1' as the default; the actual default is either a corporate character encoding or character encodings widely deployed in a certain national or regional community. For further considerations, please also see Section 5.2 of [HTML40].

7. Security Considerations

[HTML401], section B.10, notes various security issues with interpreting anchors and forms in HTML documents.

In addition, the introduction of scripting languages and interactive capabilities in HTML 4.0 introduced a number of security risks associated with the automatic execution of programs written by the sender but interpreted by the recipient. User agents executing such scripts or programs must be extremely careful to insure that untrusted software is executed in a protected environment.

8. Authors' Addresses

Daniel W. Connolly
World Wide Web Consortium (W3C)
MIT Laboratory for Computer Science
545 Technology Square
Cambridge, MA 02139, U.S.A.

EMail: connolly@w3.org
http://www.w3.org/People/Connolly/

Larry Masinter
AT&T
75 Willow Road
Menlo Park, CA 94025

EMail: LM@att.com
http://larry.masinter.net

9. References

[CLIMAPS] Seidman, J., "A Proposed Extension to HTML: Client-Side Image Maps", RFC 1980, August 1996.

[FORMDATA] Masinter, L., "Returning Values from Forms: multipart/form-data", RFC 2388, August 1998.

[HTML20] Berners-Lee, T. and D. Connolly, "Hypertext Markup Language - 2.0", RFC 1866, November 1995.

[HTML30] Raggett, D., "HyperText Markup Language Specification Version 3.0", September 1995. (Available at <http://www.w3.org/MarkUp/html3/CoverPage>).

[HTML32] Raggett, D., "HTML 3.2 Reference Specification", W3C Recomendation, January 1997.
Available at <http://www.w3.org/TR/REC-html32>.

[HTML40] Raggett, D., et al., "HTML 4.0 Specification", W3C Recommendation, December 1997.
Available at <http://www.w3.org/TR/1998/REC-html40-19980424>

[HTML401] Raggett, D., et al., "HTML 4.01 Specification", W3C Recommendation, December 1999.
Available at <http://www.w3.org/TR/html401>.

[HTTP] Gettys, J., Fielding, R., Mogul, J., Frystyk, H., Masinter, L., Leach, P. and T. Berners-Lee, "Hypertext Transfer Protocol -- HTTP/1.1", RFC 2616, June 1999.

[I18N] Yergeau, F., Nicol, G. and M. Duerst, "Internationalization of the Hypertext Markup Language", RFC 2070, January 1997.

[MHTML] Palme, J., Hotmann, A. and N. Shelness, "MIME Encapsulation of Aggregate Documents, such as HTML (MHTML)", RFC 2557, March 1999.

[MIME] Freed, N. and N. Borenstein, "Multipurpose Internet Mail Extensions (MIME) Part Two: Media Types", RFC 2046, November 1996.

[TABLES] Raggett, D., "HTML Tables", RFC 1942, May 1996.

[UPLOAD] Nebel, E. and L. Masinter, "Form-based File Upload in HTML", RFC 1867, November 1995.

[URI] Berners-Lee, T., Fielding, R. and L. Masinter, "Uniform Resource Identifiers (URI): Generic Syntax", RFC 2396, August 1998.

[XHTML1] "XHTML 1.0: The Extensible HyperText Markup Language: A Reformulation of HTML 4 in XML 1.0", W3C Recommendation, January 2000. Available at <http://www.w3.org/TR/xhtml1>.

10. Full Copyright Statement

Acknowledgement

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