Network Working Group
Request for Comments: 4155
Category: Informational

E. Hall

September 2005

# The application/mbox Media Type

## Status of This Memo



## Copyright Notice



## Abstract

This memo requests that the application/mbox media type be authorized for allocation by the IESG, according to the terms specified in RFC 2048. This memo also defines a default format for the mbox database, which must be supported by all conformant implementations.

## 1. Background and Overview

UNIX-like operating systems have historically made widespread use of "mbox" database files for a variety of local email purposes. In the common case, mbox files store linear sequences of one or more electronic mail messages, with local email clients treating the database as a logical folder of email messages. mbox databases are also used by a variety of other messaging tools, such as mailing list management programs, archiving and filtering utilities, messaging servers, and other related applications. In recent years, mbox databases have also become common on a large number of non-UNIX computing platforms, for similar kinds of purposes.

The increased pervasiveness of these files has led to an increased demand for a standardized, network-wide interchange of these files as discrete database objects. In turn, this dictates a need for a general media type definition for mbox files, which is the subject and purpose of this memo.

## 2. About the mbox Database

The mbox database format is not documented in an authoritative specification, but instead exists as a well-known output format that is anecdotally documented, or which is only authoritatively documented for a specific platform or tool.

mbox databases typically contain a linear sequence of electronic mail messages. Each message begins with a separator line that identifies the message sender, and also identifies the date and time at which the message was received by the final recipient (either the last-hop system in the transfer path, or the system which serves as the recipient's mailstore). Each message is typically terminated by an empty line. The end of the database is usually recognized by either the absence of any additional data, or by the presence of an explicit end-of-file marker.

The structure of the separator lines vary across implementations, but usually contain the exact character sequence of "From", followed by a single Space character (0x20), an email address of some kind, another Space character, a timestamp sequence of some kind, and an end-of-line marker. However, due to the lack of any authoritative specification, each of these attributes are known to vary widely across implementations. For example, the email address can reflect any addressing syntax that has ever been used on any messaging system in all of history (specifically including address forms that are not compatible with Internet messages, as defined by RFC 2822 [RFC2822]). Similarly, the timestamp sequences can also vary according to system output, while the end-of-line sequences will often reflect platform-specific requirements. Different data formats can even appear within a single database as a result of multiple mbox files being concatenated together, or because a single file was accessed by multiple messaging clients, each of which has used its own syntax for the separator line.

Message data within mbox databases often reflects site-specific peculiarities. For example, it is entirely possible for the message body or headers in an mbox database to contain untagged eight-bit character data that implicitly reflects a site-specific default language or locale, or that reflects local defaults for timestamps and email addresses; none of this data is widely portable beyond the local scope. Similarly, message data can also contain unencoded eight-bit binary data, or can use encoding formats that represent a specific platform (e.g., BINHEX or UUENCODE sequences).

Many implementations are also known to escape message body lines that begin with the character sequence of "From ", so as to prevent confusion with overly-liberal parsers that do not search for full separator lines. In the common case, a leading Greater-Than symbol (0x3E) is used for this purpose (with "From " becoming ">From "). However, other implementations are known not to escape such lines unless they are immediately preceded by a blank line or if they also appear to contain an email address and a timestamp. Other implementations are also known to perform secondary escapes against these lines if they are already escaped or quoted, while others ignore these mechanisms altogether.

A comprehensive description of mbox database files on UNIX-like systems can be found at http://qmail.org./man/man5/mbox.html, which should be treated as mostly authoritative for those variations that are otherwise only documented in anecdotal form. However, readers are advised that many other platforms and tools make use of mbox databases, and that there are many more potential variations that can be encountered in the wild.

In order to mitigate errors that may arise from such vagaries, this specification defines a "format" parameter to the application/mbox media type declaration, which can be used to identify the specific kind of mbox database that is being transferred. Furthermore, this specification defines a "default" database format which MUST be supported by implementations that claim to be compliant with this specification, and which is to be used as the implicit format for undeclared application/mbox data objects. Additional format types are to be defined in subsequent specifications. Messaging systems that receive an mbox database with an unknown format parameter value SHOULD treat the data as an opaque binary object, as if the data had been declared as application/octet-stream

Refer to Appendix A for a description of the default mbox format.

Note that RFC 2046 [RFC2046] defines the multipart/digest media type for transferring platform-independent message files. Because that specification defines a set of neutral and strict formatting rules, the multipart/digest media type already facilitates highly-predictable transfer and conversion operations; as such, implementers are strongly encouraged to support and use that media type where possible.

## 3. Prerequisites and Terminology

Readers of this document are expected to be familiar with the specification for MIME [RFC2045] and MIME-type registrations [RFC2048].

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 [RFC2119].

## 4. The application/mbox Media Type Registration

This section provides the media type registration application (as per [RFC2048]).

MIME media type name: application

MIME subtype name: mbox

Required parameters: none

Optional parameters: The "format" parameter identifies the format of the mbox database and the messages contained therein. The default value for the "format" parameter is "default", and refers to the formatting rules defined in Appendix A of this memo. mbox databases that do not have a "format" parameter SHOULD be interpreted as having the implicit "format" value of "default". mbox databases that have an unknown value for the "format" parameter SHOULD be treated as opaque data objects, as if the media type had been specified as application/octet-stream. Additional values for the format parameter are to be defined in subsequent specifications, and registered with IANA.

Encoding considerations: If an email client receives an mbox database as a message attachment, and then stores that attachment within a local mbox database, the contents of the two database files may become irreversibly intermingled, such that both databases are rendered unrecognizable. In order to avoid these collisions, messaging systems that support this specification MUST encode an mbox database (or at a minimum, the separator lines) with non-transparent transfer encoding (such as BASE64 or Quoted-Printable) whenever an application/mbox object is transferred via messaging protocols. Other transfer services are generally encouraged to adopt similar encoding strategies in order to allow for any subsequent retransmission that might occur, but this is not a requirement. Implementers should also be prepared to encode mbox data locally if non-compliant data is received.

Security considerations: mbox data is passive, and does not generally represent a unique or new security threat. However, there is risk in sharing any kind of data, because unintentional information may be exposed, and this risk certainly applies to mbox data as well.

Interoperability considerations: Due to the lack of a single authoritative specification for mbox databases, there are a large number of variations between database formats (refer to the introduction text for common examples), and it is expected that non-conformant data will be erroneously tagged or exchanged. Although the "default" format specified in this memo does not allow for these kinds of vagaries, prior negotiation or agreement between humans may sometimes be needed.

Published specification: see Appendix A.

Applications that use this media type: hundreds of messaging products make use of the mbox database format, in one form or another.

Magic number(s): mbox database files can be recognized by having a leading character sequence of "From", followed by a single Space character (0x20), followed by additional printable character data (refer to the description in Appendix A for details). However, implementers are cautioned that all such files will not be compliant with all of the formatting rules, therefore implementers should treat these files with an appropriate amount of circumspection.

File extension(s): mbox database files sometimes have an ".mbox" extension, but this is not required nor expected. As with magic numbers, implementers should avoid reflexive assumptions about the contents of such files.

Macintosh File Type Code(s): None are known to be common.

Person & email address to contact for further information: Eric A. Hall (ehall@ntrg.com)

Intended usage: COMMON

## 5. Security Considerations

See the discussion in section 4.

## 6. IANA Considerations

The IANA has registered the application/mbox media type in the MIME registry, using the application provided in section 4 above.

Furthermore, IANA has established and will maintain a registry of values for the "format" parameter as described in this memo. The first registration is the "default" value, using the description provided in Appendix A. Subsequent values for the "format" parameter MUST be accompanied by some form of recognizable, complete, and legitimate specification, such as an IESG-approved specification, or some kind of authoritative vendor documentation.

## 7. Normative 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.

[RFC2048] Freed, N., Klensin, J., and J. Postel, "Multipurpose Internet Mail Extensions (MIME) Part Four: Registration Procedures", BCP 13, RFC 2048, November 1996.

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

[RFC2822] Resnick, P., "Internet Message Format", RFC 2822, April 2001.

## Appendix A. The "default" mbox Database Format

In order to improve interoperability among messaging systems, this memo defines a "default" mbox database format, which MUST be supported by all implementations that claim to be compliant with this specification.

The "default" mbox database format uses a linear sequence of Internet messages, with each message being immediately prefaced by a separator line, and being terminated by an empty line. More specifically:

- Each message within the database MUST follow the syntax and formatting rules defined in RFC 2822 [RFC2822] and its related specifications, with the exception that the canonical mbox database MUST use a single Line-Feed character (0x0A) as the end-of-line sequence, and MUST NOT use a Carriage-Return/Line-Feed pair (NB: this requirement only applies to the canonical mbox database as transferred, and does not override any other specifications). This usage represents the most common historical representation of the mbox database format, and allows for the least amount of conversion.

- Messages within the default mbox database MUST consist of seven-bit characters within an eight-bit stream. Eight-bit data within the stream MUST be converted to a seven-bit form (using appropriate, standardized encoding) and appropriately tagged (with the correct header fields) before the database is transferred.

- Message headers and data in the default mbox database MUST be fully-qualified, as per the relevant specification(s). For example, email addresses in the various header fields MUST have legitimate domain names (as per RFC 2822), while extended characters and encodings MUST be specified in the appropriate location (as per the appropriate MIME specifications), and so forth.

- Each message in the mbox database MUST be immediately preceded by a single separator line, which MUST conform to the following syntax:

  The exact character sequence of "From";

  a single Space character (0x20);

  the email address of the message sender (as obtained from the message envelope or other authoritative source), conformant with the "addr-spec" syntax from RFC 2822;

a single Space character;

a timestamp indicating the UTC date and time when the message was originally received, conformant with the syntax of the traditional UNIX 'ctime' output sans timezone (note that the use of UTC precludes the need for a timezone indicator);

an end-of-line marker.

- Each message in the database MUST be terminated by an empty line, containing a single end-of-line marker.

Note that the first message in an mbox database will only be prefaced by a separator line, while every other message will begin with two end-of-line sequences (one at the end of the message itself, and another to mark the end of the message within the mbox database file stream) and a separator line (marking the new message). The end of the database is implicitly reached when no more message data or separator lines are found.

Also note that this specification does not prescribe any escape syntax for message body lines that begin with the character sequence of "From ". Recipient systems are expected to parse full separator lines as they are documented above.

## Author's Address

Eric A. Hall

EMail: ehall@ntrg.com

## Full Copyright Statement



## Intellectual Property



## Acknowledgement

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