Sustainability of Digital Formats: Planning for Library of Congress Collections

MBOX Email Format

>> Back

Table of Contents

Format Description Properties

- ID: fdd000383
- Short name: MBOX
- Content categories: text, email
- Format Category: file-format
- Other facets: unitary, binary, structured, symbolic
- Last significant FDD update: 2016-11-17
- Draft status: Full

Identification and description

Full name	MBOX Email Format
Description	MBOX (sometimes known as Berkeley format) is a generic term for a family of related file formats used for storing collections of electronic mail messages. MBOX formats store all of the messages of an entire folder (not an entire mailbox) in a single database file and new messages are appended to the end of the file. Each message is immediately prefaced by a separation line and terminated by an empty line. Only the the first message in an MBOX database file 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 file is implicitly reached when no more message data or separator lines are found. A message encoded in MBOX format begins with a "From " line, continues with a series of non-"From " lines, and ends with a blank line. A "From " line means any line in the message or header that begins with the five characters 'F', 'r', 'o', 'm', and ' ' (space). The "From " line structure is *From sender date moreinfo:* • *sender*, usually the envelope sender of the message (e.g., sender@sender.com) is one word without spaces or tabs • *date* is the delivery date of the message which always contains exactly 24 characters in Standard C asctime format (i.e. in English, with the redundant weekday, and without timezone information) • *moreinfo* is optional and it may contain arbitrary information. After the "From " line is the message itself in RFC 5322 format. The final line is a completely blank line with no spaces or tabs. There are four variants of MBOX: MBOXO, MBOXRD, MBOXCL and MBOXCL2. The four versions all build on the common MBOX structure and are differentiated primarily by changes to the "From " line and and the use of the "Content Length:" field in the message header in determining the start of a new message within the aggregated file. Moreover, the versions and tool sets for one version are not necessary compatible with one another. See General section for incompatibility details. MBOX files also include the message attachments, if any, in their original MIME format.
Relationship to other formats	
Defined via	IMF, Internet Mail Format
Has subtype	MBOXO, MBOXO Email Format
Has subtype	MBOXRD, MBOXRD Email Format
Has subtype	MBOXCL, MBOXCL Email Format
Has subtype	MBOXCL2, MBOXCL2 Email Format

Local use

LC experience or existing holdings	
LC preference	

Sustainability factors

Disclosure	There is no authoritative specification aside from RFC 4155. Its subtypes are partially documented and there is variation within the

	subtypes.
Documentation	Information available from a number of sources including RFC 4155, and Qmail.org.
Adoption	Prom reports that, while not a native format for many proprietary clients, MBOX (and EML) has "achieved a certain status as de facto standards because most modern email clients and servers can import and export one or both of the formats" including Thunderbird, Apple Mail, Outlook and Eudora. In addition, external programs such as Aid4Mail, Emailchemy and Xena can convert between the two formats and numerous proprietary formats. Once in an MBOX or EML format, the data can be parsed into XML using standardized schemas such as the Email Account Schema defined in the CERP project. The Smithsonian Institution Archives uses the CERP-developed toolset to normalize messages to MBOX before converting to XML. The ePADD project developed at Stanford University Libraries ingests only MBOX or IMAP accounts. Native or normalized MBOX files also can be used as access copies because they can be imported into a variety of email clients.
Licensing and patents	[Unknown, probably none].
Transparency	Text processing tools can be readily used on the plain text files used to store the email messages.
Self-documentation	The message structure helps declare the subtype but there's a lot of variation even within the established patterns.
External dependencies	None
Technical protection considerations	None

Quality and functionality factors

File type signifiers and format identifiers

Tag	Value	Note
Filename extension	mbox	MBOX database files sometimes have an "mbox" extension, but according to the specification, this is not required nor expected.
Internet Media Type	application/mbox	Not registered in IANA but listed in RFC 4155
Magic numbers	Not applicable.	MBOX database files, which are the focus of this document, do not have a magic number. As described in RFC 4155, 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. Gary Kessler states that MBOX TOC files, which act as an index to the MBOX database file, have the magic number 00 0D BB A0, followed by

		four bytes which appear to be the number of e-mails in the associated MBOX file. Comments welcome.
Pronom PUID	fmt/720	See http://www.nationalarchives.gov.uk/PRONOM/fmt/720.
Wikidata Title ID	Q285972	See https://www.wikidata.org/wiki/Q285972

Notes

General	Jonathan de Boyne Pollard describes the many incompatibilities among the MBOX formats: • Messages cannot be reliably read from MBOXO and MBOXRD format mailboxes by MBOXCL and MBOXCL2 readers. • There is no guarantee that any "Content-Length:" headers in the original message are correct and appropriate, which are preserved exactly as they are by MBOXO and MBOXRD. • MBOXCL2 readers cannot return messages with no "Content-Length:" headers. • Messages cannot be reliably read from MBOXCL2 format mailboxes by MBOXO or MBOXRD readers. • Delivering messages to MBOXCL2 format mailboxes with MBOXO or MBOXRD tools will corrupt the mailbox, rendering all subsequently delivered messages irretrievable • Because "From " at the start of a line is more probable than ">From " in real-world messages, an MBOXRD reader will restore a greater number of messages written to a mailbox by an MBOXO tool to their original forms than an MBOXRD tool, but will not restore all messages. • Conversely, when an MBOXO reader is used, less message corruption will be observed in the final results if the messages were written by an MBOXO tool than if they were written by an MBOXRD tool. Wikipedia reports that "different MBOX formats use various mutually incompatible mechanisms to enable message file locking, including fcntl(), lockf(), and "dot locking" which are problematic in network mounted file systems, such as the Network File System (NFS). Because more than one message is stored in a single file, some form of file locking is needed to avoid the corruption that can result from two or more processes modifying the mailbox simultaneously. This could happen if a network email delivery program delivers a new message at the same time as a mail reader is deleting an existing message. MBOX files should be locked also while they are being read. Otherwise the reader may see corrupted message contents if another process is modifying the mbox at the same time, even though no actual file corruption occurs." Because MBOX stores the contents of an entire folder in one file, the size of the MBOX single file can become exceedingly large. Any

	corruption in the file may affect the ability of certain clients to access individual messages or even the entire folder.
History	The naming scheme was developed by Daniel J. Bernstein, Rahul Dhesi, and others in 1996. Each version originated from a different version of Unix.

Format specifications

- RFC 4155: The application/mbox Media Type (http://tools.ietf.org/rfc/rfc4155.txt). The application/mbox Media Type
- mbox - file containing mail messages (http://qmail.org/qmail-manual-html/man5/mbox.html). This is the de facto standards document for MBOX formats
- See IMF

Useful references

URLs

- MBOX is a family of several mutually incompatible mailbox formats (https://jdebp.eu/FGA/mail-mbox-formats.html). Useful history and discussion by Jonathan de Boyne Pollard
- Importing and exporting your mail (http://kb.mozillazine.org/Importing_and_exporting_your_mail#Mbox_files). Information about adoption and importing/exporting MBOX variants
- WiscMail / WiscMail Plus - Importing "mbox" mailboxes into desktop email clients (https://kb.wisc.edu/wiscmail/page.php?id=6436).
- Mail Summary Files (http://www.jwz.org/doc/mailsum.html). From 2000, use of MBOX variants in Netscape Mail
- Using Maildir Format (http://cr.yp.to/proto/maildir.html). Describes the Maildir format which became an alternative to MBOX
- Wikipedia: MBOX (http://en.wikipedia.org/wiki/Mbox).
- Preserving Email by Christopher J. Prom. DPC Technology Watch Report 11-01 (http://dx.doi.org/10.7207/twr11-01). Very helpful discussion of the challenges in preserving email collections, especially in describing workflows and tools.
- CERP: The Collaborative Electronic Records Project (http://siarchives.si.edu/cerp/). Takes email accounts that have been normalized to MBOX files and parses them out to XML following an email preservation schema.
- MUSE (http://mobisocial.stanford.edu/muse/). A tool for working with email archives developed by Stanford's Mobisocial Lab
- Preservation of Electronic Mail Collaboration Initiative (http://www.history.ncdcr.gov/SHRAB/ar/emailpreservation/reports.htm). Collaborative effort to develop XML schema for email records
- Tool sets for MBOX
  - XENA (http://xena.sourceforge.net/). Email migration and conversion tool developed by the National Archives of Australia
  - EmailChemy (http://www.weirdkid.com/products/emailchemy/). Proprietary Java application tool for email format migration and conversion tool
  - Aid4Mail (http://www.aid4mail.com/). Desktop email migration and conversion tool
  - MailStore (http://www.mailstore.com/en/mailstore-home.aspx). Conversion tool, proprietary but free for personal use

- PRONOM entry for fmt/720 (http://nationalarchives.gov.uk/PRONOM/fmt/720). Information in PRONOM from UK National Archives about MBOX family. PUID: fmt/720
- The mbox Format: How email clients store mail on your hard disk (http://email.about.com/cs/standards/a/mbox_format.htm).
- WikiData entry for Q285972 (https://www.wikidata.org/wiki/Q285972). Information in WikiData about MBOX family. WikiData Title ID: Q285972.

Last Updated: 09/23/2019

Digital Preservation Home | Digital Formats Home