www.photocase.de

Archiving Email Messages with Hypermail

CLEANING UP

Hypermail converts email messages to HTML and allows you to group your messages in tidy archives. BY ANDREA MÜLLER

Nearly everyone keeps postcards and letters from their loved ones. And even if you don't, you probably keep at least one file at home with letters from authorities, banks, and insurance companies. If your important documents are filed in an organized and accessible way, you are more likely to find a document you need when you come back later. Why not apply the same principle to old email messages, using Hypermail *http://www.hypermail.org*?

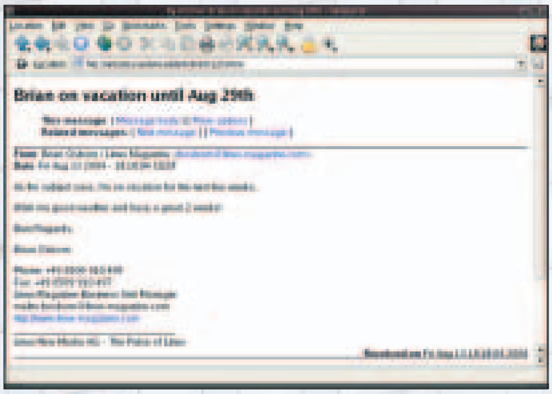
Hypermail converts your email messages into HTML documents (Figure 1). Each document contains links to any preceding or answering messages in the thread. The program stores attachments in a subfolder and places a link to the attachment in the HTML file. To allow you to find the messages you are looking for, Hypermail also generates a number of index pages, where the messages are sorted by subject, author (Figure 2), date, and thread. Additionally, Hypermail generates an *attachment.html* file with a list of mail attachments.

Package or Home-Grown?

Some distributions – such as Suse Linux Professional, Mandrake Power Pack, or Debian – include Hypermail by default. If Hypermail is included with the distribution, you can simply run your distribution's package manager to install it. If your distribution does not include Hypermail, or if you would prefer to use the latest version of the mail archiver, you will need the source code to build Hypermail yourself. The source archive is available from the project homepage. Hypermail is unlikely to give you a hard time. Assuming that the *gcc*, *make*, and *glibc-devel* packages are available on your system, you can simply type *./configure*, *make* and *su -c "make install"* to build and install the program. The call to *make install* copies the program and accompanying files to a directory below */usr/local*.

A Question of Formats

Hypermail only accepts the mbox format as input (see the box titled "mbox Format"). Some mail programs – Evolution and Mozilla, for example – use the mbox file format by default to store email messages. If you use one of these applications, all you need to do is to create a separate folder for the files you want to archive. The mbox file will have the same name as the folder in the mail client. Users with Mozilla need to look at the directory tree below their profile folder *~/.mozilla/default/xxxxxxx/Mail*, where *xxxxxxx* is an arbitrary string that Mozilla uses to identify the profile. You should find a directory with the same name as your mail account; this is where Mozilla stores mbox files without a filename extension. Evolution stores mail in a folder called *~/.evolution*.



Brian on vacation until Aug 29th

Figure 1: Messages archived by Hypermail include the message body, a selection of headers, and links to the previous and following messages in the discussion thread.

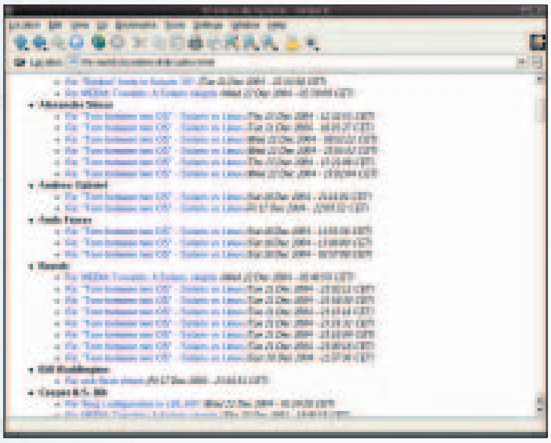


Figure 2: Hypermail generates an index that sorts your mail alphabetically by author.

If your mail client does not use mbox format, it may still have a function for creating an mbox folder, or it may allow you to export messages into mbox format. KMail asks you to specify the format when you create a new folder (Figure 3) and stores the folder below the *~/Mail* directory. Sylpheed users can use the *Export to Mbox file* function in the *File* menu.

Quickstart

After installing Hypermail, enter the following command

```
hypermail -m mailbox -d ➥
outputdirectory
```

to create your first mail archive. Specify the path to your mbox file with the *-m* parameter. *-d* is the directory where Hypermail will create the archive. You do not need to create the directory before running the command; Hypermail will take care of that automatically. When the program has finished, you should discover the HTML-formatted messages and the index file in the target directory. The *index.html* file contains a thread list by default; you can click on one of the links to go to a message. If you like, you can move to one of the other indexes (author, date, sender or attachments).

Hypermail supports several languages other than English. If you prefer to use a language other than English for the HTML pages, add *-L*, followed by the language parameter, as follows:

```
hypermail -L es -m mailbox -d ➥
folder
```

sets the language to Spanish. Besides Spanish, Hypermail can give you Italian (*it*), Russian (*ru*), and German (*de*), for example. The *-h* flag gives you a list of supported languages in the line that starts with *-L*.

The headings on the index pages include the archive name – Hypermail sets this to the name of the mbox file. To assign an individual name, use the *-l* command line switch followed by the name. It is not necessary to run Hypermail multiple times to merge multiple mailboxes to a single HTML archive; instead, simply specify the names of the mbox files you want to merge as the *-m* parameter value. Hypermail allows you to add messages to an archive at a later date. The *-u* parameter tells Hypermail to update the archive.

```
hypermail -u -m newbox -d folder
```

adds the messages in *newbox* to an archive in the *folder* directory. Of course, Hypermail will update the index files to reflect the changes.

Tailor-Made Archives

Let's assume you want to add messages that you have composed yourself, and that are now sitting in your Outbox, to the archive, and your mail program does not generate **message IDs**. You cannot use the default setting here, as this just provokes an error message: *Message-ID is missing, ignoring message with subject 'subject'*. The *-o require_msgids=0* option tells Hypermail to process messages of this kind. *-o* is short for options, and believe me, the program has quite a few of them. You can type *man hmrc* for a list.

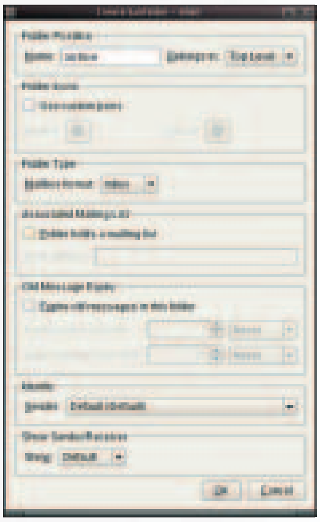


Figure 3: KMail allows you to specify the format when creating a folder.

If you are archiving mailboxes with a large number of entries, you might like to take a closer look at the *monthly-index=1* and *folder_by_date* options. The former option tells Hypermail to add an overview to the *index.html* file, which points to monthly indexes (Figure 4). This is a good thing speed-wise, as an index file for a few thousand messages can easily reach a size of 10MByte or more. In contrast to this, the browser should load a monthly index in next to no time. This option does not mean that the HTML files will be placed in separate folders, however. To distribute your files over multiple monthly folders, you need the Hypermail *folder_by_date* option. Let's combine this feature with a monthly index:

```
hypermail -m mbox -d ➥
folder -o ➥
monthly-index=1 -o ➥
folder_by_date=%y%m
```

The *%y%m* is a so-called format string, where *%y* stands for the year, and *%m* for the month when the message was created. This command tells Hypermail to create subfolders with names such as *0312* below the output directory. Messages from December 2003 would be stored in this subfolder. If you prefer to have the month first, simply

mbox Format

mbox is the generic term for a family of legacy mailbox formats used by Unix hosts to store mail messages. The name reflects the approach that *mbox* adopts for storing multiple messages in a single mailbox file. The various *mbox* formats were categorized by Daniel J. Bernstein, Rahul Dhesi, and others in 1996 and include *mboxo*, *mboxrd* , *mboxcl*, and *mboxc12*.

Within an *mbox* file, incoming messages are appended to the file using the "From " string to identify the beginning and an empty line to identify the end of each message. Because the *mbox* format was never officially specified in an RFC, it leaves programmers of mail user agents some scope for designing incompatible formats.

An example *mbox* file is shown in Listing 1.

switch the order of the format string: *folder_by_date = %m%y*

If you are planning to publish archives on a web server, for example, the *-o spamprotect = 1* option is a good idea. This option tells Hypermail to modify the mail addresses. Instead of *name@domain*, the program writes *name_at_domain*. This makes it more difficult for spammers to harvest target addresses. You can use the *-o antispam_at = replacementcharacter* option to tell Hypermail what to write instead of the @ character.

Listing 1: An *mbox* file

```
From user@example.com  Sat Jun 14 14:45:12 2003
Received: from localhost (localhost.localdomain [127.0.0.1])
        by anmen.not-for-mail (8.11.6/8.11.6) with ESMTP id h5ECjBA29295
        for
; Sat, 14 Jun 2003 14:45:11 +0200
Message-ID: <3EEB0E35.C0077C5@example.com>
Date: Sat, 14 Jun 2003 13:59:49 +0200
From: User Domain
To: a414@sedacon.com (Marc Andre Selig)
Subject: Test mail

Hello!

From a414@sedacon.com  Sat Jun 14 14:48:14 2003
Date: Sat, 14 Jun 2003 14:48:14 +0200
From: a414@sedacon.com
To: a414@sedacon.com (Marc Andre Selig)
Subject: Another test

Yet another test.
```

If the body text of a message just happens to have an empty line followed by a line starting with “From” and then a blank space, the “From” string is replaced by “>From” so the line will not look like the start of a new message.

Stringing messages together in a single large file makes for efficient use of the inodes on a filesystem. On the downside, *mbox* files become slower and less responsive as they grow. Another disadvantage of the *mbox* format is that it requires locking, so that multiple programs will not access the file in parallel.

The program has a few more useful features, such as a quote tagging option. If you want to highlight quotes, rather than just using the quote character (typically >), you could specify *-o iquotes = 1* to use italics for quotes. The *-o linkquotes = 1* option is also useful. This option tells Hypermail to generate a link from the first quote to take you to the original message.

Less Typing

Options are a good thing, but they do have a downside: nobody can remember all of them. This typically means checking the manpage every time you need to run the archiving tool. Once you have found a set of options that are perfect for your requirements, there is a clever way of avoiding having to enter them every time you run Hypermail. When Hypermail launches, it parses the *.hmrc* file in your home directory. This means you can specify the number of header lines you want to see; you can even define monthly folders and the path to the mailbox file. Options are written just like they are on the command line. Let’s look at an example with the command line switch *-o*. The following entry

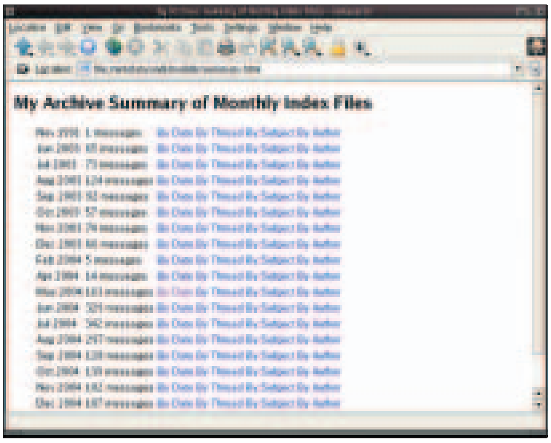
```
require_msgids=0
```

tells Hypermail to archive messages without a message ID. Command line options have priority over entries in *.hmrc*. You can specify a default mailbox but still change the mailbox using the *-m* option, followed by the input file name. Listing 2 is a sample *.hmrc*.

If you tidy up regularly, you can look forward to a quicker and less cluttered mail program. And you can browse older messages any time you like. The Hypermail indexes look great, even if you use a text-based browser. If your archives are taking up too much of your hard disk, you can simply swap an archive out to a CD. ■

GLOSSARY

Message ID: A unique number in the email header, which is comprised of an arbitrary local part, the @ character, and a domain part. The message ID could be abcdefghijkl@example.com, for example. The uniqueness of the message ID is useful for Usenet, as most news servers simply reject messages with a message ID they have seen before.



My Archive Summary of Monthly Index Files

Nov 2000 1 messages By Date By Thread By Subject By Author
Jun 2003 65 messages By Date By Thread By Subject By Author
Jul 2003 73 messages By Date By Thread By Subject By Author
Aug 2003 124 messages By Date By Thread By Subject By Author
Sep 2003 52 messages By Date By Thread By Subject By Author
Oct 2003 57 messages By Date By Thread By Subject By Author
Nov 2003 74 messages By Date By Thread By Subject By Author
Dec 2003 66 messages By Date By Thread By Subject By Author
Feb 2004 5 messages By Date By Thread By Subject By Author
Apr 2004 14 messages By Date By Thread By Subject By Author
May 2004 183 messages By Date By Thread By Subject By Author
Jun 2004 321 messages By Date By Thread By Subject By Author
Jul 2004 342 messages By Date By Thread By Subject By Author
Aug 2004 297 messages By Date By Thread By Subject By Author
Sep 2004 128 messages By Date By Thread By Subject By Author
Oct 2004 135 messages By Date By Thread By Subject By Author
Nov 2004 182 messages By Date By Thread By Subject By Author
Dec 2004 187 messages By Date By Thread By Subject By Author

Figure 4: An additional monthly index makes large archives easier to handle.

Listing 2: ~/.hmrc Options

```
#create Spanish language page
language=es

#European date format
eurodate=1

#Standard mailbox
mbox=/home/andi/archiv

#Links to quotes
linkquotes=1

#Create monthly folders
folder_by_date=%y%m

#Display headers
showheaders=1

#Header lines that Hypermail should display
show_headers=From,To,Subject,Date,Message-ID,User-Agent,X-Mailer,X-Newsreader

#Display quotes in italics
iquotes=1

#Do not archive messages with X-Hypermail-Deleted in header
deleted=X-Hypermail-Deleted

#No mailto link
mailto=none

#Create monthly index
monthly_index=1
```