splitter - a C program to split Usenet archives into separate threads
(an updated version 8/12/92 - see Recent Notes and Fixes below)

When I considered the benefits of decomposing the comp.sys.next.*whatever* feed so I could index it under Librarian, I wrote a *simple* C program to do it. When splitting the archive into filenames reflecting the subject of each thread, I also wanted to remove any offensive characters to the indexing function of Librarian. My program is called 'splitter' and requires the following syntax in a shell:

```
> splitter prefix archive1 archive2 ...
```

where prefix is a 1-6 character prefix for each created file. The prefix is requested in order to prevent undesirable first characters in filenames. Splitter was designed to operate on the *Month.Z* files (after they're uncompressed) available in the /pub/news directory on nova.cc.purdue.edu.

Installation and Testing

I have included the source. You can compile it with

```
> cc -O splitter.c -o splitter
```

Splitter always opens output files in the append modeÐi.e. previous messages with identical subjects are preserved for multiple runs of splitter. I have enclosed two archived files of old comp.sys.next (files *Aug* and *Sep 91*) which you can use to test splitter. Simply cd to the splitter directory and type

```
> splitter csn Aug Sep
```

You should see messages indicating that splitter is operating on these files, and when finished, several filenames with *csn_* (for comp.sys.next) prefixes should reside in the directory. To remove these test files type

```
> rm csn_*
```

I recommend installing a copy of splitter in each sub-directory in which you want to use splitter, since I currently don't support splitting a file in one directory and sending the output results to another. The splitter executable is only 50 kB. My directory tree for comp.sys.next is

Graphic Omitted here...See README in package

where a copy of splitter resides in each sub-directory. I have also included eight icons for the sub-directories. If you like them, copy these to the appropriate directory and rename as .dir.tiff . The icons are 60x60 pixels. They look good in Librarian, but the Workspace scales them to 48x48. The results in Librarian are shown below.

Graphic Omitted here...See README in package

How it Works

For the output files, an underscore (_) is appended after the prefix. The beginning of each message is then determined by locating the string *From:* and then the normally subsequent *Subject:* or *Subject: Re:* string. The subject line is stripped of the leading *Subject:* or *Subject: Re:* and any undesirable characters for use in a filename are replaced with an underscore (currently left brackets, right brackets, periods, blanks, and slashesÐif you avoid these characters in file and directory names you'll save yourself a bunch of headaches with Librarian).

The prefix_+modified_subject_line is then used as the filename for containing the message. Note that any replies which (exactly) reference a previous subject will be appended to an already existing file, thus preserving the discussion thread. Therefore, it makes sense to use this program to sequentially split monthly archives in order to preserve the order of posting. I suggest waiting until month's end and a completed archive before splitting, or you may easily duplicate messages.

I have used this program to do incremental additions to my indexed archives and both the splitter and the Librarian's index updating features have functioned flawlessly under NeXTstep 2.1. Also note, although not a problem for filenaming, that some filenames will possess several underscores, or trailing underscores, depending on the subject name (*. . .* is an example of a string which converts to several adjacent underscores).

Some Minor Problems

I do not promise that this program will always perform as desired, since some messages do not conform to the *From: Subject:* header. However, I have observed that messages which contain a *From:* header line and then a blank line will all be written to a file named *prefix_*, which is easily located using the Workspace browser. But, if any text immediately follows the *From:* header line, the message will be written to a filename composed of the modified first line of that text. Messages which do not have a *From:* header line will be written into the file of the previous message regardless of the *Subject:* header line.

A rare but occasional *From:* will appear in a message body at the beginning of a line, which does cause *splitter* to open a new file. If everyone would follow the convention of placing some sort of indicator in front of quoted material, this last problem would disappear.

Recent Notes and Fixes (this release)

In using splitter I have encountered a few problems:

1) When several months worth of a large archive have been split within one directory, the directory becomes extremely large and disk access seems to slow appreciably. I have worked around this by starting a new directory every 6 months or so. Ideally to get around this I could allow threads starting with some range of letters to be broken into separate directories automatically, so the bulk of a single archive would not grow to tremendous sizes. Maybe in a future release.

2) When moving between disks or to OD it is most efficient to tar and compress the entire Librarian archive before moving. Unfortunately, my previous version very often resulted in filenames which were cumulatively larger than 100 characters, which is the current limit for the version of tar provided in NeXTstep 2.1. In this version I have limited filenames to a maximum of 70 characters. This allows up to 30 more characters for the path. If this doesn't suit you, you can change the source.

3) Sometimes the *Subject:* or *Subject: Re:* headers have slightly varying spacing. I have added tests to account for several cases (see source).

Disclaimer

Use at your own risk and feel free to modify *splitter* as desired (maybe add a NeXT front end). You should be able to compile and run this source on any machine with an ANSI C compiler.

Michael McCulloch 8/12/92 (original version released 12/12/91)
Independent NeXT Developer
Huntsville, AL
(205) 726-1832
sorry, but I'm not setup for email at this time

P.S. This is way too much information to provide for a simple C program, isn't it?