NAME

bibsplit − split large BibTeX bibliography files into independent parts

SYNOPSIS

bibsplit [**−?**] [**−author**] [**−bycentury**] [**−bydecade**] [**−byhalfcentury**] [**−bylabel**]
[**−bynumber** *nnnn*] [**−bypentad**] [**−byrange** *range-list*] [**−byscore**] [**−byyear**]
[**−copyright**] [**−filter** *command*] [**−help**] [**−logfile** *filename*] [**−maxopen** *nnnn*]
[**−outfile** *filename*] [**−prefix** *xxx*] [**−quick**] [**−silent**] [**−tmpdir** *dirname*] [**−version**]
<*infile* or *bibfile1 bibfile2 bibfile3* . . .
>*outfile*

DESCRIPTION

bibsplit splits large bibliographic database files into smaller independent parts, selecting destination files according to requests made by **−by***xxx* command-line options, which allow selection by citation labels, by count of entries, and by groups of publication years.

If you want to select entries by more complex criteria, such as author names, keywords, subject classifications, title words, etc., then **bibsplit** is not the tool you need: use **bibextract**(1) instead.

As long as BibTeX is asked to retrieve only a limited number of citations from database files, it does not matter how many citations there are, or how big the database files are. BibTeX simply processes each file in sequential order, and since the files are read only once, and internal processing of string lookups uses a fast constant-time algorithm, access time is strictly proportional to the amount of data read and written.

However, it is frequently desirable to be able to typeset a complete bibliography file, so that one can verify that each entry can be correctly processed by BibTeX, and correctly typeset by TeX.

This is readily done with a simple TeX file that looks like this:

```
\input btxmac
\bibliographystyle{plain}
\nocite{*}
\bibliography{mybib}
\bye
```

or a corresponding LaTeX2e file that looks like this:

```
\documentclass{article}
\bibliographystyle{plain}
\begin{document}
  \nocite{*}
  \bibliography{mybib}
\end{document}
```

Splitting large bibliographic files is sometimes necessary, because

- internal table sizes in most TeX and BibTeX implementations limit the number of entries actually extracted from one or more BibTeX database files to about 4000;
- large database files are undesirable for World-Wide Web and FTP file transfers across slow network connections;
- large database files are slower to edit;
- large database files are more prone to massive editing disasters.

bibsplit provides the needed solution to this problem, and does so with at most two or three passes over the input data.

In the first pass, **bibsplit** writes temporary files containing all non-*@String* entries, partitioned according to the command-line options chosen. It saves all *@String* definitions in memory, and it builds up a list in memory of which definitions are needed by each of the temporary files.

In the second pass, **bibsplit** writes the required *@String* definitions into the final files, sorted in ascending lexicographic order, followed by the contents of their corresponding temporary files, and then deletes the temporary files. No further parsing is needed for the second pass, so it is relatively fast.

For user feedback, **bibsplit** writes a brief progress report to *stdout* at important stages of its work. If you do not want to see this, then simply redirect *stdout* to the null device: on UNIX, **bibsplit . . . >** */dev/null* or else use the **−silent** option.

At the time of writing, **bibsplit** processes BIBTEX data at about 1MB/sec on a fast modern UNIX workstation, so practical applications should never take more than a few seconds.

OPTIONS

Command-line options may be abbreviated to a unique leading prefix, and letter case is ignored.

To avoid confusion with options, if a filename begins with a hyphen, it must be disguised by a leading absolute or relative directory path, e.g., */tmp/-foo.bib* or *./-foo.bib*.

GNU- and POSIX-style options of the form **− −name** are also recognized: they begin with two option prefix characters.

In the event of conflicting **−by***xxx* options, the last one specified takes precedence.

−? Display brief usage information on *stderr* and exit with a success status code before processing any input files.

This is a synonym for **−help**.

−author Show author information on *stderr* and exit with a success status code before processing any input files.

−bycentury Split the input bibliography stream into output files suffixed by a four-digit century.

−bydecade Split the input bibliography stream into output files suffixed by a four-digit year identifying the starting year of the decade.

−byhalfcentury Split the input bibliography stream into output files suffixed by a four-digit year identifying the starting year of the half century.

−bylabel Split the input bibliography stream into 26 output files suffixed by a single lowercase letter identifying the initial letter of the citation labels.

This option is equivalent to, and shorthand for,
−byrange a,b,c,d,e,f,g,h,i,j,k,l,m,n,o,p,q,r,s,t,u,v,w,x,y,z.

−bynumber *nnnn* Split the input bibliography stream into output files with no more than *nnnn* non-*@String* entries.

As a special case, a zero number is interpreted as infinity; see the end of this section for a practical application.

This option creates output files suffixed by a four-digit entry count reflecting the input order of the first entry in that file, and entries are written to those output files in strict input order.

−bypentad Split the input bibliography stream into output files suffixed by a four-digit year identifying the starting year of the pentad (a five-year interval).

−byrange *a-e,f-h,i-l,m-p,q-r,s-w,x-z*
Split the input bibliography stream into output files suffixed by a letter range taken from the comma-separated range list, identifying the initial letters of the citation labels. Ranges can be indicated by either hyphen or underscore, and that character will also be used in the output BIBTEX file names.

The list shown above is only illustrative; you can
choose any sensible letter grouping.

Lettercase is ignored in the range list.

Use this option when you want coarser grouping, and larger output files, than provided by **−bylabel**.

−byscore Split the input bibliography stream into output files suffixed by a four-digit year identifying the starting year of each score (twenty) of years.

−byyear Split the input bibliography stream into output files suffixed by a four-digit year, one year per output file.

−copyright Show copyright information on *stderr* and exit with a success status code before processing any input files.

−filter *command* On completion of splitting, apply *command* to each output split file, producing a temporary output file, and if that succeeds, replace the original output split file by that temporary file. If *command* fails, silently delete the temporary file.

If *command* contains spaces or other characters that are significant to the shell, then of course it needs to be surrounded by protecting quotes, or the special characters need to be prefixed by a backslash. **bibsplit** will surround *command* with apostrophes (single quotes), so they cannot be used in *command*. Should you require apostrophes, then you must embed your commands inside a short executable script file, and use that for *command*.

This option is most useful for applying **bibsort**(1) to the output files, because even if the input bibliography was already sorted, resolution of citations and cross-references will have destroyed that order.

−help Display brief usage information on *stderr* and exit with a success status code before processing any input files.

This is a synonym for **−?**.

−logfile *filename* Redirect warning and error messages from *stderr* to the indicated filename. This option is provided for user convenience on poorly-designed operating systems (e.g., IBM PC DOS) that fail to provide for redirection of *stderr* to a specified file.

This option can also be used for discarding messages, with, e.g., on UNIX systems, **−logfile** */dev/null.*

−maxopen *nnnn* All operating systems have limits, sometimes Draconian, on the number of simultaneously open files, and **bibsplit**, particularly with the **−bylabel** or **−byyear** options, may hit them.

To avoid the need for multiple applications of **bibsplit**, this option limits the number of simultaneously open files to *nnnn*. This does not increase the number of passes made over the input stream, but may cause additional file closing and opening.

On most modern UNIX systems, and in real applications, this option should rarely be needed.

Benchmarks show no noticeable effect on runtime when small values of *nnnn* are chosen, but because **bibsplit**'s implementation language offers no way to test for an open-file limit-exceeded condition, and because that limit varies between operating systems and installations, and on some, even depends on other current user processes and resource quotas, no sensible default value for *nnnn* can be chosen that is guaranteed to work everywhere.

−outfile *filename* Redirect output from *stdout* to the indicated filename. This option is provided for user convenience on operating systems that fail to provide for redirection of *stdout* to a specified file.

bibsplit uses *stdout* only for a brief progress report, so there is never much data written to it.

−prefix *xxx* Supply a prefix for the output file names. If this option is omitted, then the basename of the current input filename (*including* any leading directory path) is used. When no input filename is available, then *stdin* is used.

The suffixes attached to the output filenames contain no leading separator character, so, for example, the command **bibsplit −byscore** *gnats.bib* for a bibliography containing entries from 1941 to 2001 would produce output files *gnats1940.bib*, *gnats1960.bib*, *gnats1980.bib*, and *gnats2000.bib*.

If you prefer a separator, do it like this: **bibsplit −byscore −prefix** *gnats- gnats.bib* to get output files named *gnats-1940.bib*, etc.

The **−prefix** *xxx* option may include a directory path, so **bibsplit −byscore −prefix** */usr/tmp/gnats- gnats.bib* would write the split files in the directory */usr/tmp*.

In the interests of maximal filename portability, **bibsplit** assumes that slash, backslash, and colon are directory component separators, and legal characters in filenames are letters, digits, hyphen, underscore, and dot; all others will be removed.

−quick Suppress reading of the initialization files, *$LIBDIR/.bibsplitrc*, *$HOME/.bibsplitrc*, and *./.bibsplitrc*. *LIBDIR* represents the name of the **bibsplit** installation directory; it is not a user-definable environment variable.

Normally, the contents of those files, if they exist, are implicitly inserted at the beginning of the command line, with comments removed and newlines replaced by spaces. Thus, those files can contain any **bibsplit** options defined in this documentation, either one option, or option/value pair, per line, or with multiple options per line. Empty lines, and lines that begin with optional whitespace followed by a sharp (#) are comment lines that are discarded.

If the initialization file contains backslashes, they must be doubled because the text is interpreted by the shell before **bibsplit** sees it.

−silent Suppress output of progress reports to *stdout*.

−tmpdir *dirname* Use the file directory *dirname* for temporary files. Otherwise, following the common UNIX practice, **bibsplit** will use the directory specified by the environment variable *TMPDIR*, or if that is not set, then */tmp*.

This option may also be spelled **−tempdir**.

−version Show version information on *stderr* and exit with a success status code before processing any input files.

If no **−by***xxx* options are given, **bibsplit** defaults to **−bynumber** *2000*, producing a split into files about half the maximum practical size, with ample room for future additions.

In the event that bibliography entries are encountered that cannot be assigned to a suitable output file according to the particular **−by***xxx* option chosen (or assumed by default), they will be written to a file whose basename is suffixed by the uppercase string *UNKNOWN*.

Similarly, *@String* definitions that are not used in any input bibliography entry will not be written to the normal split files, so they are collected, sorted, and written to a separate file whose basename is suffixed by the uppercase string *UNUSED*. The basename of that file is determined by that of the *last* input BibTeX file.

You could use this feature to find and remove unused *@String* definitions, like this:

bibsplit -bynumber *0 mybib.bib*
mv *mybib.bib mybib.bib-old*
mv *mybib-000001.bib mybib.bib*

If a duplicate *@String* definition is encountered, then a warning is issued if the definitions differ, except possibly at whitespace. Multiple differing definitions are collected and later output together in the same order they were read, so as not to lose information.

COMMENT HANDLING

The original BibTeX specification did not have a clearly-defined comment syntax, but the BibTeX grammar defined by the author in the lengthy article *Bibliography prettyprinting and syntax checking*, TUGboat,

14(3) 222--222, **14**(4) 395--419, (1993) does: as in TEX, comments begin with percent, and run to end-of-line. That article is included in the **bibclean**(1) distribution.

bibclean(1) and **bibsplit** assume that an input file takes the form

```
% FILE HEADER COMMENTS
@Preamble{...}
% preamble comments
@Preamble{...}
% preamble comments
@Preamble{...}
...
% STRING BLOCK COMMENTS
@String{...}
% string comments
@String{...}
% string comments
@String{...}
...
% ENTRY BLOCK COMMENTS
@Book{...}
% entry comment
@Article{...}
% entry comment
@TechReport{...}
...
% FILE TRAILER COMMENTS
```

Blank or empty lines may appear anywhere, and are thus not shown in this sketch.

This organization has been found to be the most useful in many hundreds of BIBTEX files containing hundreds of thousands of document entries, and is very similar to that commonly found in well-written computer software for several decades. In particular, comments always *precede* the code or data that they refer to; they never follow.

Any of the comment regions may be empty, and after the *@String* definitions, BIBTEX entries for any supported document type may appear, and in any order.

The comment blocks in UPPERCASE take precedence over other comment blocks, and will be transferred verbatim to *every* output file containing BIBTEX entries, preserving the order shown above.

All other comments are assumed to refer to the nearest following *@Name{...}* group, and will be attached to those groups, and output when they are output, preserving that order.

All input lines that are blank or empty are discarded. However, for readability and editing convenience, **bibsplit** takes care to incorporate blank lines around all bibliographic entries, just as **bibclean**(1) does.

Any other text which does not conform to the BIBTEX grammar is converted to a comment, and thus preserved in at least one of the output files.

CROSS-REFERENCE HANDLING

In most cases, BIBTEX file entries are completely independent of one another, except for use of abbreviations from *@String* definitions, which **bibsplit** already handles nicely.

However, in some types of bibliographies, entries use the BIBTEX *crossref = "label"* facility to include additional data from another entry; the commonest such case is an *@InProceedings* entry that cross references a following *@Proceedings* entry.

Sometimes, an entry may contain a note with a citation of another entry, such as an article series where part I cites part II, or an article citing a subsequent erratum. For closely-related articles, it is useful to include such citations in the BIBTEX files, so that an author who remembers to cite only one of a series of related publications will automatically get bibliography entries for all of them.

In both these cases, the cross-referenced or cited entry follows the one that references it, so **bibsplit**, after reading the original entry, is able to examine it, and prepare a list of entries that it refers to. When **bibsplit** later encounters those entries, it outputs them not only to their normal split file, but also to all of the other files that contain earlier entries that refer to them.

Backward references are, however, more challenging. For example, an article erratum might contain a citation of the original paper, which appears earlier in a bibliography ordered by publication time. In this case, **bibsplit** will have already output the original entry without knowing that it will later be cited, and because it makes no attempt to hold all entries in memory (a strategy that would routinely fail on small systems), the cross reference has arrived too late for it to act. BIBTEX itself would require a following LATEX or TEX run to deposit the citation into the auxiliary file, in which a second BIBTEX run would find it, and finally correctly incorporate the cross reference in the typeset bibliography.

To deal with this important case of backward references, while still being frugal with memory, **bibsplit** takes a different approach. As each entry is output to a split file, **bibsplit** augments a list with entries (citation-label, BIBTEX-filename), so that it knows in which file each entry has been written. It also records the citation labels of any embedded references in a to-be-found list, with entries of the form (citation-label, BIBTEX-filename-1, BIBTEX-filename-2, . . ., BIBTEX-filename-*n*). It then looks in the to-be-found list to see if this entry is needed by earlier entries already written to other files as well, and if so, outputs it to those.

On completion of processing of all of the input stream, and generation of any unused-labels file, it then reexamines the to-be-found list, sorts it by filename, and then steps through these files in order, reading entries, and writing each one found out to all of the BIBTEX file(s) in which it has been referenced, but does not yet appear. Any citation label from the to-be-found list which is not in the original (citation-label, BIBTEX-filename) list is diagnosed as an unsatisfied reference, since its absence is definitely an error in the bibliography.

In the worst case, this algorithm will result in **bibsplit**'s reading the input data a total of three times, but in most cases, only a few of the split files, and sometimes, none, need to be read again.

CAVEATS

BIBTEX has loose syntactical requirements that the current simple implementation of **bibsplit** does not support. In particular, outer parentheses may *not* be used in place of braces following ''@keyword'' patterns. If you have such a file, you can use **bibclean**(1) to prettyprint it into a form that **bibsplit** can handle successfully.

Because all *@String* definitions are saved in memory, and all citation labels as well, very large jobs may exceed the memory requirements of very small systems. About two megabytes of memory should suffice for the vast majority of practical applications.

ENVIRONMENT VARIABLES

TMPDIR Name of directory where **bibsplit** writes its temporary files. Its value is ignored if a command-line **−tmpdir** *dirname* option is given. [default: */tmp*]

FILES

In the following, *LIBDIR* represents the name of the **bibsplit** installation directory; it is not a user-definable environment variable. If **bibsplit** has been installed properly at your site, the value of *LIBDIR* is

```
/usr/local/share/lib/bibsplit/bibsplit-1.00
```

`$LIBDIR/.bibsplitrc` System-specific initialization file containing customized **bibsplit** command-line options.

`$HOME/.bibsplitrc` User-specific initialization file containing customized **bibsplit** command-line options.

`./.bibsplitrc` Current-directory-specific initialization file containing customized **bibsplit** command-line options.

`$LIBDIR/bibsplit.awk` **awk**(1) program invoked by **bibsplit**.

SEE ALSO

awk(1), **bawk**(1), **bibcheck**(1), **bibclean**(1), **bibdup**(1), **bibextract**(1), **bibjoin**(1), **biblabel**(1), **biblex**(1), **biborder**(1), **bibparse**(1), **bibsearch**(1), **bibsort**(1), **bibtex**(1), **bibunlex**(1), **bstpretty**(1), **citesub**(1), **emacs**(1), **gawk**(1), **lacheck**(1), **latex**(1), **mawk**(1), **nawk**(1), **tex**(1).

AUTHOR

Nelson H. F. Beebe
Center for Scientific Computing
University of Utah
Department of Mathematics, 322 INSCC
155 S 1400 E RM 233
Salt Lake City, UT 84112-0090
USA
Email: beebe@math.utah.edu, beebe@acm.org, beebe@ieee.org (Internet)
WWW URL: http://www.math.utah.edu/~beebe
Telephone: +1 801 581 5254
FAX: +1 801 585 1640, +1 801 581 4148

AVAILABILITY

bibsplit is freely available; its master distribution can be found at

```
ftp://ftp.math.utah.edu/pub/tex/bib/
http://www.math.utah.edu/pub/tex/bib/
```

in the file *bibsplit-x.yy.tar.gz* where *x.yy* is the current version. Other distribution formats are usually available at the same location.

That site is mirrored to several other Internet archives, so you may also be able to find it elsewhere on the Internet; try searching for the string *bibsplit* at one or more of the popular Web search sites, such as

```
http://altavista.digital.com/
http://search.microsoft.com/us/default.asp
http://www.dejanews.com/
http://www.dogpile.com/index.html
http://www.euroseek.net/page?ifl=uk
http://www.excite.com/
http://www.go2net.com/search.html
http://www.google.com/
http://www.hotbot.com/
http://www.infoseek.com/
http://www.inktomi.com/
http://www.lycos.com/
http://www.northernlight.com/
http://www.snap.com/
http://www.stpt.com/
http://www.yahoo.com/
```

COPYRIGHT