

NAME

`biblabel` – generate standardized BibTeX citation labels

SYNOPSIS

```
biblabel [ --? ] [ --author ] [ --copyright ]
  [ --corporate-file input-words-to-ignore-file ] [ --dump-corporate-file output-file ]
  [ --dump-ignore-file output-file ] [ --dump-label-file output-file ] [ --help ]
  [ --ignore-file input-words-to-ignore-file ] [ --long-corporate-names ]
  [ --used-file input-labels-in-use-file ] [ --version ]
  [BIBTeXfile(s)] >outpub-citesub-file
```

DESCRIPTION

biblabel filters a BIB_TE_X bibliography, or bibliography fragment, on its standard input, or one or more bibliographies named on the command line, and prints on standard output, *stdout*, lines containing pairs of old and new citation labels, suitable for input to the companion program, **citesub**(1).

The citation label is formed by these rules, easily applicable by a human, or by a computer program like this one:

- (1) Take the first author's last name, dropping apostrophes, Jr/Sr/generation numbers, protecting braces, and eliminating accents (e.g., $J\{\backslash"a\}nsch \rightarrow Jaensch$, and $Jind\{v\{r\}ich \rightarrow Jindrich$), using multi-letter transliterations if that is conventional. Preserve hyphenated names, like Baeza-Yates, in full.
- (2) Append a colon.
- (3) Append the four-digit year of publication.
- (4) Append another colon.
- (5) Pick the initial letters of at most *three* of the leading important words in the title that begin with a letter, excluding articles, prepositions, and \TeX math mode, and append those letters.

For example, given the title *{Euler}'s Constant to \$1271\$ Places*, (from an article by Donald E. Knuth in *Mathematics of Computation*, **16**(79) 275--281, July 1962) this recipe produces *ECP*.

- (6) If the resulting citation label is already in use, add a letter *a*, *b*, *c*, ... to make it unique. In those rare cases when there are more than 26 such collisions, add additional letters, producing suffixes written in a base-26 number system in ascending order: *a*.*.z*, *aa*.*.az*, *ba*.*.bz*, ..., *za*.*.zz*, *aaa*.*.aaz*, ..., *zza*.*.zzz*, *aaaa*.*.aaaz*, ..., *zzza*.*.zzzz*, ...

This will produce a label like `Smith:1994:ABC`.

The reason for including a four-digit year is that the worldwide Y2K problem at the millennium change amply demonstrated the foolishness of two-digit year abbreviations. Also, some bibliographies may be historical, with entries dating back hundreds of years. Using a four-digit year will keep sorts of otherwise identical keys in chronological order, and putting the year before the key derived from the title will facilitate sorting by year by, e.g., **bibsort**(1).

Because any change in citation labels must be accompanied by a change in citations in all documents that use the bibliography, it is not sufficient to just produce a new bibliography file with changed labels. Consequently, the output of **biblabel** is expected to be saved, and subsequently used with **citesub**(1) to actually carry out the substitutions efficiently. If no documents other than the bibliography file itself need to be changed, then a simple UNIX or IBM PC DOS pipeline of the form

```
biblabel <foo.bib | citesub -f -foo.bib >foo.new
```

will produce a new bibliography file with all of the citation labels changed to the new standardized form.

To avoid confusion between labels with common prefixes, such as *Smith80* and *Smith80a*, **citesub**(1) will check for leading context of a left brace, quote, comma, whitespace, or beginning of line and trailing context of a right brace, comma, quote, percent, whitespace, or end of line so as to match these styles:

```

@Book{Smith:1980:ABC,
crossref = "Smith:1980:ABC",
crossref = {Smith:1980:ABC},
\cite{Smith:1980:ABC}

\cite{Smith:1980:ABC,Jones:1994:DEF}

\cite{%
  Smith:1980:ABC,%
  Jones:1994:DEF%
}

```

Created labels are guaranteed to be unique within the input files provided on the command line.

However, in a larger project, one may wish to exclude labels that are already in use in other bibliographies. To provide for this, the **--used-file** option can be specified to define the name of a file of labels that are already in use.

When a label is found in use, and the current file matches the in-use label filename, the label is considered to be unused; otherwise, repeated runs through this program would keep changing already-assigned labels.

OPTIONS

Unless otherwise noted, all command-line options may be abbreviated to a unique leading prefix, and letter case is *not* significant.

All options are parsed before any input bibliography files are read, no matter what their order on the command line.

The leading hyphen that distinguishes an option from a filename may be doubled, for compatibility with GNU and POSIX conventions. Thus, **-author** and **--author** are equivalent.

Except for the options described below, all other command-line words are assumed to be input files. Should such a filename begin with a hyphen, it must be disguised by a leading absolute or relative directory path, e.g., */tmp/-foo.bib* or *./-foo.bib*.

--author Display an author credit on *stderr*, and then terminate with a success return code (0 on UNIX).

--copyright Display a copyright statement on *stderr*, and then terminate with a success return code (0 on UNIX).

This option may be abbreviated **--c**.

--corporate-file *input-words-to-ignore-file*

Specify the name of a file containing additional words (one or more per line) to be added to the internal corporate name ignore list, after converting to lowercase and stripping nonletters. Such words, and all words in the normal ignore list, are not considered when corporate name abbreviations are constructed.

Multiple **--corporate-file** options may be specified.

See also the **INITIALIZATION FILES** section below.

--dump-corporate-file *output-file*

Dump the corporate name ignore list on the specified file, and then terminate with a success return code. This file is valid for subsequent use with the **--corporate-file** option.

To avoid disastrous overwriting of a bibliography file in the event of a command-line mistake, the output file must not yet exist.

- dump-label-file** *output-file*

Dump the labels-in-use list on the specified file. This file is valid for subsequent use with the **--used-file** option.

Unlike the other dump options, which are processed at startup time, this option is processed only at the end of a successful execution.

To avoid disastrous overwriting of a bibliography file in the event of a command-line mistake, the output file must not yet exist.
- dump-ignore-file** *output-file*

Dump the ignore list on the specified file, and then terminate with a success return code. This file is valid for subsequent use with the **--ignore-file** option.

To avoid disastrous overwriting of a bibliography file in the event of a command-line mistake, the output file must not yet exist.
- f** *labels-in-use-file*

This option is deprecated, but retained for compatibility with versions of **biblabel** prior to version 0.04. It may be removed in a later version. See instead the **--used-file** description below.
- help** or **--?**

Display a brief help message on *stderr*, and then terminate execution immediately with a success return code (0 on UNIX).
- ignore-file** *words-to-ignore-file*

Specify the name of a file containing additional words (one or more per line) to be added to the internal ignore list, after converting to lowercase and stripping nonletters. Such words are ignored when the up-to-three-letter title abbreviations, and corporate name abbreviations, are constructed.

Multiple **--ignore-file** options may be specified.

See also the **INITIALIZATION FILES** section below.
- long-corporate-names**

Do not abbreviate corporate names.

Without this option, a braced corporate author/editor string of *{Free Software Foundation}* is reduced to *FSF*. With this option, it becomes *Free-Software-Foundation*.

Single-word corporate names are never abbreviated to an initial: *IBM* remains that way, instead of being reduced to *I*.
- used-file** *labels-in-use-file*

Specify the name of a file containing citation labels already in use.

Each line consists of a whitespace-separated pair of filename and citation label. Inclusion of the filename in which the label is already in use is required, both so that it can be used in diagnostic messages, and to avoid unnecessary changes to labels in the current file.

This option can be used in a multi-file bibliography collection to guarantee unique citation labels across the entire collection.

Multiple **--used-file** options may be specified.

See also the **INITIALIZATION FILES** section below.
- version**

Display a program version number and date on *stderr*, and then terminate with a success return code (0 on UNIX).

INITIALIZATION FILES

To make it easy to have personal customized lists of words to be ignored, and labels in use, on startup, **biblabel** looks for default initialization files, and then processes any additional files specified by command-

line **--corporate-file**, **--ignore-file**, and **--used-file** options.

These files may contain:

- Comments, which run from a percent (%) or sharp (#) character anywhere on the line to the end of that line. Comments are stripped *before* any further processing.
- Empty lines containing zero or more space and/or tab characters. Such lines are ignored.
- Data lines.

The three default optional initialization files are:

biblabel.cig List of words to add to the internal corporate name ignore list, after converting to lowercase and stripping nonletters.

biblabel.ign List of words to add to the internal ignore list, after converting to lowercase and stripping nonletters.

biblabel.use List of citation labels already in use, together with the names of the files in which they are used.

Each line consists of a whitespace-separated pair of filename and citation label.

If these files exist, they must be in the current directory.

To make it possible to override the built-in ignore list, if the special word **@RESET@** appears in any ignore-list file, then it, and all entries in the internal list, are immediately forgotten.

For consistency, this special word is also recognized in labels-in-use files, but has limited utility since there is no built-in citation label table. It could nevertheless be useful if **biblabel** were wrapped inside another script, or a shell alias, which themselves provided command-line initialization files.

INTERNAL DEFAULTS

biblabel starts with a built-in ignore list, containing lowercase forms of words that are to be ignored when forming the up-to-three-letter title abbreviations, and corporate name abbreviations. It then augments that built-in list with the contents of any *biblabel.ign* file in the current directory, plus any files specified with **--ignore-file** options.

Here is the built-in ignore list, taken directly from the output file created by the **--dump-ignore-file** option. To conserve space here, the original one-word-per-line list has been reformatted into paragraphs of words with common initial letters.

```
%% Title: Dump of ignore list
%% CreationDate: Fri Mar  9 10:16:31 MST 2001
%% Creator: biblabel version 0.04 [06-Mar-2001]
%% For: Nelson H. F. Beebe <beebe@suncore.math.utah.edu>
%% Directory: /u/sy/beebe/tex/biblabel/biblabel-0.04

a ab aber als also am an and any are as at auf aus aux
away az

be bei bin bir bist but by

cum

da dans das dat de dei dem den der des det di die dos
down

e een eene egy ei ein eine einen einer eines eit el en
er es et ett eyn eyne
```

```

for from fuer fur

gehabt gl gli

ha hab habe haben habt had haette hai has hast hat
hatte have he heis hen hena henas het hin hinar hinir
hinn hith ho hoi

i il in into is ist its

ka ke

l la las le les lo los

mia mit

n na ne nicht nji not

o oben oder of off ohne on onto or os others out over

pas

s seid sie sind so sur

t ta that the these this those to

uber um uma un una und une uno unter unto up

via vom von

with without

y yr

zu zum zur

```

biblabel also has a much shorter built-in list of words to be ignored when forming corporate name abbreviations. The list is small, because all words from the main ignore list are also excluded when forming such abbreviations. **biblabel** augments the built-in list with the contents of any *biblabel.cig* file in the current directory, plus any files specified with **--corporate-file** options.

Here is the built-in corporate ignore list, taken directly from the output file created by the **--dump-corporate-file** option. To conserve space here, the original one-word-per-line list has been reformatted into paragraphs of words with common initial letters.

```

%% Title: Dump of corporate ignore list
%% CreationDate: Fri Mar  9 10:22:54 MST 2001
%% Creator: biblabel version 0.04 [06-Mar-2001]
%% For: Nelson H. F. Beebe <beebe@suncore.math.utah.edu>
%% Directory: /u/sy/beebe/tex/biblabel/biblabel-0.04

co company corp corporation

gmbh group

```

inc incorporated

limited ltd

staff

team

There are no internal defaults for the list of citation labels already in use. That list is initialized from any *biblabel.use* file in the current directory, plus any files specified with `--used-file` options.

DIAGNOSTIC MESSAGES

biblabel issues diagnostic messages on *stderr* in the standard form *filename:linenumber:message-text*. When the program is run under an advanced text editor like **emacs**(1) or **vim**(1), the editor can then jump to the error location with only a keystroke or two.

Here, in alphabetical order, are the messages that **biblabel** can produce, with brief explanations. Each is prefixed with an uppercase word denoting the severity level.

ERROR: duplicate citation label [xxx]

This condition will definitely cause erroneous substitutions when **citesub**(1) is used, so the error *must* be corrected and **biblabel** run again on the corrected file.

NOTE: generational suffixes are normally not preceded by a comma: [xxx]

See the next diagnostic description.

NOTE: Junior-like suffixes are normally preceded by a comma: [xxx]

Although a few authors (e.g., noted computer scientist Guy L. Steele Jr.) omit the separating comma before the suffix, it is usually a coding mistake in the **BIBTEX** file.

This practice may be changing in American English: the widely-followed *Chicago Manual of Style*, 14th ed., University of Chicago Press (Chicago and London), 1993, ISBN 0-226-10389-7, notes in section 8.55 on p. 307:

Traditionally, Jr. and Sr. have been set off with commas, whereas I, II, III, IV, and so on have not. This tradition is still widely followed, and the University of Chicago Press recognizes and accepts it; but the Press now also accepts, and in fact recommends, that the commas be omitted in both cases.

On the other hand, the *MLA Handbook for Writers of Research Papers*, 4th ed., Modern Language Association of America (New York), 1996, ISBN 0-87352-565-5, in section 4.6.1 on p. 110, wants a comma before all such suffixes.

Strunk and White's *The Elements of Style* 3rd ed., Macmillan (New York), 1979, ISBNs 0-02-418230-3 and 0-02-418220-6, on p. 3 says to omit the comma before *Jr.*, but to include it before *Ph.D.* and *S.J.* suffixes.

Words into Type, 3rd ed., Prentice Hall (Englewood Cliffs, NJ), 1974, ISBN 0-13-964262-5, recommends the comma before *Jr.* and *Sr.*, but notes that newspapers frequently omit it.

The *ACS Style Guide*, 2nd ed., American Chemical Society (Washington, DC), 1997, ISBNs 0-8412-3461-2 and 0-8412-3462-0, requires a comma before *Jr.* and *Sr.*, but says to treat roman numeral suffixes according to the person's preference.

In other words, style manuals don't agree! **biblabel**'s diagnostic is thus only informational.

PORTABILITY: unexpected 8-bit character(s) found in author/editor name [xxx]

For maximum portability across systems, **BIBTEX** files should not use characters outside of the 7-bit ASCII character set, since standard **T_EX** control sequences can represent all needed accents. Even at sites where extended 8-bit encodings are used in **T_EX** and **L_AT_EX** files for convenience and improved hyphenation, those files can remain site-specific, while the bibliography database can be shared worldwide.

This message cannot be raised unless the underlying **awk**(1) implementation supports 8-bit characters in

regular-expression patterns. **gawk**(1) and **mawk**(1) do, but others may not.

WARNING: incomplete accent removal [xxx] -> [yyy]

biblabel contains many heuristic reductions of accented letters to unaccented ones, but occasionally an accent combination is met that it cannot recognize, and the result is that the generated label may not be exactly what was intended. Only rarely will manual editing of the output substitution file be necessary.

WARNING: unexpected standalone Junior-like name suffix [xxx]

A name suffix (Jr., Sr., III, ...) was found where a full name was expected. This can happen when simple-minded software tools are used to convert bibliographic data to $\text{BIB}\text{T}\text{E}\text{X}$ form. This message would be raised if *P. D. Q. {Bach, Jr.}* or *Bach, P. D. Q., Jr.* had been incorrectly coded as *P. D. Q. Bach and Jr.*

WARNING: unrecognized text [xxx]

There is a syntax error in the *biblabel.use* file, or a file specified by the **--used-file** option. Such files are expected to contain comments (from percent (%) or sharp (#) to end-of-line), empty or blank lines, and lines with whitespace-separated pairs of filename and citation label.

WARNING: year [xxxx] out of acceptable range [1000..2099] for citation label: using 20xx instead

A valid year is required for generation of a citation label; an out-of-range, or omitted, year will be represented by *20xx* in the generated citation label.

SEE ALSO

bibcheck(1), **bibclean**(1), **bibextract**(1), **bibdestringify**(1), **bibdub**(1), **bibextract**(1), **bibjoin**(1), **biblex**(1), **biborder**(1), **bibparse**(1), **bibsearch**(1), **bibsort**(1), **bibsplit**(1), **bibtex**(1), **bibunlex**(1), **cite-sub**(1), **emacs**(1), **vim**(1).

AUTHOR

Nelson H. F. Beebe
 University of Utah
 Department of Mathematics, 110 LCB
 155 S 1400 E RM 233
 Salt Lake City, UT 84112-0090
 USA
 Email: beebe@math.utah.edu, beebe@acm.org,
 beebe@computer.org (Internet)
 WWW URL: <http://www.math.utah.edu/~beebe>
 Telephone: +1 801 581 5254
 FAX: +1 801 581 4148

AVAILABILITY

biblabel 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/index-table-b.html#biblabel>

in the files

biblabel-x.yy.jar
biblabel-x.yy.shar.gz
biblabel-x.yy.tar.gz
biblabel-x.yy.zip
biblabel-x.yy.zoo

where *x.yy* is the current version. Each of the popular archive format unpacks into an identical distribution tree in a subdirectory named *biblabel-x.yy*. [**Caution:** older software distributions may omit the leading subdirectory prefix in some archive formats.]

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 *biblabel* at one or more of the popular Web search sites, such as

```

http://search.microsoft.com/
http://www.altavista.com/
http://www.dejanews.com/
http://www.dogpile.com/
http://www.euroseek.net/
http://www.excite.com/
http://www.go2net.com/
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.websmostlinked.com/
http://www.yahoo.com/

```

COPYRIGHT

```

#####
#####
#####
###
###      biblabel: generate standardized BibTeX citation labels      ###
###
###      Copyright (C) 1994, 1996, 1997, 2001, 2006 Nelson H. F. Beebe  ###
###
###      This program is covered by the GNU General Public License (GPL), ###
###      version 2 or later, available as the file COPYING in the program ###
###      source distribution, and on the Internet at                      ###
###
###      ftp://ftp.gnu.org/gnu/GPL                                       ###
###
###      http://www.gnu.org/copyleft/gpl.html                            ###
###
###      This program is free software; you can redistribute it and/or   ###
###      modify it under the terms of the GNU General Public License as  ###
###      published by the Free Software Foundation; either version 2 of  ###
###      the License, or (at your option) any later version.            ###
###
###      This program is distributed in the hope that it will be useful,  ###
###      but WITHOUT ANY WARRANTY; without even the implied warranty of  ###
###      MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.  See the  ###
###      GNU General Public License for more details.                    ###
###
###      You should have received a copy of the GNU General Public       ###
###      License along with this program; if not, write to the Free     ###
###      Software Foundation, Inc., 59 Temple Place, Suite 330, Boston,  ###
###      MA 02111-1307 USA                                              ###
#####
#####
#####

```