NAME

bibjoin − join duplicate or similar entries in a BibTeX bibliography file

SYNOPSIS

bibjoin [**−author**] [**−check-missing**] [**−copyleft**] [**−copyright**] [**−ignore-characters** *regexp*] [**−keep-duplicate-values**] [**−version**] [BibTeXfile(s) or < infile] > outfile

DESCRIPTION

bibjoin filters one or more BIBTEX bibliographies, or bibliography fragments, from the specified files, or from its standard input if no filenames are provided, printing on standard output a bibliography in which adjacent duplicate, or similar, entries have been joined into one entry. Such action may be necessary when bibliography entries are collected from many sources.

bibjoin should be applied to a bibliography file only after entries have been suitably ordered so that candidates for joining appear consecutively. This can be done mostly automatically if standardized citation labels are first generated, perhaps by **biblabel**(1) and **citesub**(1), or by the GNU **emacs**(1) *bibtex-insert-standard-BibNet-citation-label* function from the *bibtools* library, then the bibliography is sorted by citation labels, such as by **bibsort**(1).

Only a human reader can reliably decide when two bibliography entries are truly the same. **bibjoin** can help automate much of this work, but manual editing will almost certainly still be necessary. If two entries are joined, these conditions must be satisfied:

- identical citation labels;
- identical year;
- if CODENs are given in both entries, the CODEN lists must be identical;
- if ISBNs are given in both entries, the ISBN lists must be identical;
- if ISSNs are given in both entries, the ISSN lists must be identical;
- if a journal article entry, identical volume, and if both have page numbers, identical initial page numbers.

An empty value, or a value containing only space and/or question marks, is equivalent to an omitted value for the purposes of these comparisons. The reason for this choice is that question marks have proved to be useful indicators of *unknown* values, distinguished from *omitted* values.

When two ‘equal’ value strings are found for the same key, one of them is normally deleted. Otherwise, both key/value pairs are output. Manual editing will then be required to choose between them.

Special handling is supplied for ‘author’ and ‘editor’ fields. When a personal name appears in two forms, one with initials, and one without, such as ‘P. D. Q. Bach’ and ‘Philippe D. Q. Bach’, the names are considered to match, and the longer form is retained. In addition, to deal with the UnCover database practice of omitting authors 3, 4, ..., N-1, two author/editor personal name lists are considered to match if one has 3 names and the other more than 3, and the first, second, and last match as above; the longer form is retained.

Special handling is supplied for ‘bibdate’ fields, provided they are in either of the forms

Wed Jul 6 15:27:50 1994
Wed Jul 6 15:27:50 MDT 1994

If either of the values is unrecognized, then separate key/value pairs are preserved. Otherwise, only the more recent of the two dates is kept.

Special handling is supplied for ‘pages’ entries. If entries are found with identical initial page numbers, but one of them has question marks in place of the final page number, or has no final page number at all, such as "123--127", "123--??", and "123", then the ones with the question marks or no final page numbers will be dropped. This facilitates merging in data from library databases that do not record final page numbers.

Value strings are considered equal if they match after all characters other than letters, digits, and plus are removed, and letter case is ignored. (The default set of retained characters can be redefined via the **−ignore-characters** *regexp* option described later.) For ‘title’ entries, leading words ‘A’, ‘An’, ‘On’, and ‘The’ are ignored, because some library databases drop them. Value strings are also considered to match if

one is an exact prefix of the other, because truncation of author lists and titles is a common problem in journal databases. This fuzzy equality helps to eliminate many match failures that arise from minor variations in punctuation, spacing, and capitalization. **bibjoin** has no way of determining which of the two strings should be preserved, so it uniformly discards the shorter one (which presumably has less 'information'): this choice will frequently be *wrong*! The shorter string will be preserved if the **−keep-duplicate-values** option described later is used.

If two *title* or *booktitle* strings have the same length, and match when letter case is ignored, then the one with more capitalized words is saved. The reason for this choice is that some library databases arbitrarily downcase titles, losing information that should be preserved.

Syntax errors in the input stream will cause abrupt termination with a fatal error message and a non-zero exit code. The output will be incomplete, so you should always examine the output file before assuming that you can replace the input file with the output file.

If the **−keep-duplicate-values** option has been specified, then key/value pairs in output entries are sorted alphabetically by key name, so that duplicate keys arising from the join operation appear consecutively, simplifying the subsequent manual editing task. Otherwise, keys are ordered according to the conventions of **biborder**(1).

After completion of manual corrections, it is recommended that the bibliography be processed by **biborder**(1) to standardize key/value order (if the **−keep-duplicate-values** option was used), and to check for any remaining duplicate keys or citation labels.

OPTIONS

Command-line options may be abbreviated to a unique leading prefix. 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.

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*.

−author Print author information on *stderr* and exit immediately with a successful status code.

−check-missing If this option is specified, missing expected key fields will be supplied, with the key field name prefixed with OPT, and the value string set to a pair of question marks, e.g.

```
OPTvolume =    "??",
```

The *OPT* prefix ensures that the key is ignored by BibTeX, so that the question marks will not appear in an output *.bbl* file. The GNU Emacs *bibtex-mode* editing support has functions for removing the OPT prefixes, and so does **bibclean**(1).

The doubled question marks are distinguished from single ones that might legitimately appear in value strings, and also serve as a convenient regular-expression pattern for **bibextract**(1), allowing easy preparation of a printed listing of just those entries that have incomplete bibliographic data:

bibextract *" '[?][?]' BibTeXfiles* **| lpr**

−copyleft Print copyright information on *stderr* and exit immediately with a successful status code.

−copyright Print copyright information on *stderr* and exit immediately with a successful status code.

−ignore-characters *regexp* Specify a regular expression to define the set of characters to be ignored in value string comparisons. The default is *'[^A-Za-z0-9+]'*.

−keep-duplicate-values Instead of discarding the shorter of two value strings that are considered 'equal', preserve the shorter of them using the key suffixed with the letter 'z', e.g., *title* and *titlez*. If such a key already exists, add additional suffixing 'z'

letters to make the key unique.

−version Display the **bibjoin** version number and date on *stderr* and exit immediately with a successful status code.

WARNING AND ERROR MESSAGES

bibjoin will issue warning messages in the following cases:

- With **−check-missing**, for unrecognized BIBTEX entry types. The entry will be output without checking for missing key names.
- For duplicate key names. Such key/value pairs are sorted together by name, preserving their original order.
- When identical key/value pairs are reduced to a single pair by discarding duplicates.

bibjoin will issue an error message and terminate with exit code 1, and *incomplete output*, in the following cases:

- for an unrecognized command-line argument (only the minimal unique prefix of each option is currently examined);
- end-of-file is reached while collecting an entry or value;
- a line beginning with '@' is encountered while collecting an entry, before balanced braces have been found.

CAVEATS

BIBTEX has loose syntactical requirements that the current simple implementation of **bibjoin** 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 **bibjoin** can handle successfully.

SEE ALSO

bibcheck(1), **bibclean**(1), **bibdup**(1), **bibextract**(1), **biblabel**(1), **biblex**(1), **biborder**(1), **bibparse**(1), **bibsearch**(1), **bibsort**(1), **bibtex**(1), **bibunlex**(1), **citesub**(1), **emacs**(1).

AUTHOR

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