

NAME

`bibsearch` – search BibTeX bibliography files

SYNOPSIS

```
bibsearch [ --bibsearchpath dirname[:dirname:dirname:...] ] [ --? ] [ --debug ] [ --help ]
[ --quiet ] [ --version ] [ database ]
```

DESCRIPTION

bibsearch uses the **mgquery**(1) database search engine to provide superfast searching in one of several collections of BIB_TE_X bibliography database files. The most active databases are generally updated nightly. **bibsearch** normally displays the date of the last update, plus some statistics on the size of the collection. It also gives some helpful hints about commonly-used commands.

Multiple databases are available, and others may be added from time to time without updating this documentation, or the installed **bibsearch** software; use the **--help** option occasionally to find out what databases are currently installed.

The default database at the **bibsearch** development site at the University of Utah Mathematics Department is **tug**, the largest, and most active. The default may be different at your site; use the **--help** option to find out.

At the time of writing, these databases are available:

dblp Universität Trier, Germany, Digital Bibliography and Library Project (DBLP) bibliography archive. Coverage of computer science journals, conferences, and technical reports, with more than 187,000 entries [**updated irregularly**].

karlsruhe Universität Karlsruhe, Germany, combined world-wide Computer Science bibliography archive, with more than 965,000 entries [**updated weekly**]. This collection includes a major part of the **tug** archive described below, but its file mirror may often be weeks or months behind.

mathutah Mathematics Department, University of Utah, research groups bibliographies, with about 200,000 entries [**updated nightly**].

tug BibNet Project and T_EX User Group combined bibliography archive, with more than 279,000 entries [**updated nightly**]. As of early 2001, the **tug** collection includes about 200 journal-specific bibliographies, of which about 70 provide *complete* coverage of their respective journals.

The archives covers these subject areas in computer science, electronic document representation, and mathematics:

- ANSI, IEC, IEEE, Internet, and ISO computer-related standards;
- *American Mathematical Monthly* and *Canadian Journal of Mathematics* == *Journal canadien de mathématiques*;
- computer architectures (Intel IA-64 and visual instruction sets);
- computer arithmetic;
- computer graphics and visualization;
- computer science conferences and journals (*hundreds* of them);
- cryptanalysis, cryptography, cryptohistory, cryptology, and encryption;
- database research;
- electronic document representation;
- fonts;
- GNU Project of the Free Software Foundation;
- HTML and SGML, and other SGML-based markup languages, such as ChemML, MathML, MusicML, VRML, and XML;

- Internet;
- *Lecture Notes in Computer Science* and *Lecture Notes in Computational Science and Engineering*;
- literate programming;
- mathematical physics (one journal only);
- numerical linear algebra;
- operating systems (GNU/Linux, Mach, Plan9, UNIX);
- page representation languages (POSTSCRIPT and PDF);
- programming languages (Axiom, BMDP, Fortran, Icon, Java, Macsyma, Maple, Mathematica, POSTSCRIPT, Python, Reduce, SAS, S-Plus, SPSS);
- quantum chemistry (one journal, one book series, and one personal bibliography, only);
- statistics;
- supercomputing;
- symbolic algebra;
- T_EX and Metafont;
- typography and typesetting;
- Unicode; and
- the X Window System.

OPTIONS

All options can be abbreviated to a unique leading prefix, and either the normal UNIX single-hyphen prefix, or the GNU/POSIX double-hyphen prefix, may be used.

--bibsearchpath *dirname[:dirname:dirname:...]*

Provide a search path of one or more directory trees in which to find **mg**(1) databases, overriding the normal installation default of `/usr/local/src/bibdata`. Each directory in the search path contains a separate subdirectory for each database, and each such subdirectory contains an `mgdata/bibfiles` subdirectory with the actual files used by **mgquery**(1).

Multiple **--bibsearchpath** options accumulate, to avoid long ugly colon-separated directory lists.

Since a particular user will often want to specify the same search path each time **bibsearch** is run, the path can also be set as the value of the **BIBSEARCHPATH** environment variable. However, command-line use of **--bibsearchpath** overrides any **BIBSEARCHPATH** setting.

As a special convenience, an empty directory path element, represented by a leading or trailing colon, or two adjacent colons between directory names, is replaced by the local system default **bibsearch** directory search path. This makes it easy for users to *augment*, rather than *override*, the search path, and to do so *without having to know what the default search path is*.

Thus, a typical user setting might be

```
$HOME/bib:
```

to put the user's own databases ahead of the system ones, overriding any of those with the same names as the user's databases, or

```
:$HOME/bib
```

to put them behind the system ones, preventing hiding of the system databases.

It is possible for the directory path to include directories to which the current user does not have read access; such directories will be silently ignored. This feature is intentional: it makes it possible for individual users, or groups, to register their databases with the local **bibsearch** installer, who need only augment the **DEFAULTBIBSEARCHPATH** setting in the installed **bibsearch** program, so that they automatically see them when they run **bibsearch**. However, by suitable setting of directory permissions, they can control access by others, without having to involve local system management.

This feature is not necessarily antisocial and unfriendly. Bibliographic data is largely entered by humans, so it is highly error prone. A user might wish to have personal **bibsearch** access to rough bibliographic data that has not yet been polished up enough for public view. Once that cleanup happens, a simple directory permission change can instantly release the data to others on the local system.

- `--?` Same as `--help`. Since the query character is significant to UNIX shells, you have to suitably quote it: write `--'?', --"?",` or `--\?`.
- `--debug` Turn on debug tracing, showing the directory path search operation, and the **mgquery** (1) environment and invocation. Debug messages are written to *stderr*.
Although it should rarely be necessary, you can also define the **DEBUG** environment variable to a nonempty string to turn on debugging. This lets it take effect a little earlier than if it were requested on the command line.
- `--help` Give a brief help message on *stdout*, show the system, user, and expanded directory search paths, list the available databases, and exit with a success return code (0 on UNIX).
- `--quiet` Suppress the display of the usual verbose startup banner that offers brief instructions for **mgquery**(1) search techniques.
- `--version` Display the program version number and data on *stdout*, and exit with a success return code (0 on UNIX).
- database* Specify the database to use, overriding the default. Use the `--help` option to get a list of available databases, with brief descriptions, and to show the system, user, and expanded directory search paths.
At the **bibsearch** development site, this list includes at least these: **dblp**, **karlsruhe**, **mathutah**, and **tug**.

SEARCHING

The **mg**(1) search engine in **bibsearch** is very simple to use: almost anything that you type is treated as text to search for, and the search begins when you end your input line.

mg(1) commands begin with an initial period, and are discussed further below.

Each search result normally begins with a separator line of dashes, followed by a World-Wide Web *Universal Resource Locator* (URL) address that uniquely identifies the location of the bibliography file. Here is an example, from a search for *Knuth typesetting book*:

```
----- 65786

URL = ftp://ftp.math.utah.edu/pub/tex/bib/font.bib Line=1453

KEYWORDS = display generation printing publishing

@Book{Knuth:1979:TMN,
  author =      "D. E. Knuth",
  title =      "{TEX} and {METAFONT}, New Directions in
                Typesetting",
  publisher =   "Digital Press",
  address =    "Billerica, MA",
  pages =      "360",
```

```

year =          "1979",
ISBN =          "0-932376-02-9",
LCCN =          "Z253.3 .K58 1979",
bibdate =      "Tue May 12 10:13:36 1998",
bibsource =    "Graphics/imager/imager.books.bib,
Graphics/siggraph/79.bib,
ftp://ftp.math.utah.edu/pub/tex/bib/siggraph/new/79.bib",
annote =       "A landmark book at the time it was
published. Newer versions exist. Less than
portable as claimed, but still significant.
Required reading for anyone doing font
design and type setting.",
keywords =     "general references, standards text books,
software, programming systems, character
display/generation, Applications,
printing/publishing industry, general
references, standards text books and
software, programming systems, character
display/generation and Applications,
printing/publishing industry",
}

```

----- 65934

Each output BIB_TE_X entry is surrounded by blank lines (a *paragraph* in T_EX) for better visibility, and to facilitate easy text selection in GNU **emacs**(1): put the cursor anywhere in the entry, and type M-h (*mark-paragraph*) then M-w (*kill-ring-save*) to copy the text into the editor and window system paste buffers. Text processing languages, like **awk**(1), **icon**(1), **perl**(1), and **ruby**(1), and search utilities like **agrep**(1) and **glimpse**(1), also have simple mechanisms for dealing with paragraph-sized chunks of text.

Searches always *ignore* letter case.

Searching is based on **words**, where a **word** is a consecutive string of *letters*, *digits*, *hyphen* (*dash* or *minus*), *underscore*, *backslash*, or *apostrophe*. This permits searching for ordinary words, as well as for ISSN and ISBN values, for programming language variable names, for T_EX control sequences, and for names like “O’Reilly”.

All other characters are treated as word separators, so to search for “*input/output*”, you must search for two words, “*input*” and “*output*”. Similarly, you could search for an e-mail address “*rms@gnu.org*” with the string “*rms gnu org*”, assuming the default query-ranked search mode.

Partial word matches are not usually accepted: if you search for “*tex*”, neither “*text*” nor “*texture*” will match. However, **mg**(1) will ‘stem’ search words, removing common English suffixes, so a search for “*mathematical*” will first be reduced to “*mathemat*” and that will match “*mathematical*”, “*mathematics*”, and “*mathematician*”. Regrettably, the current version of **mgquery**(1) does not provide any way to suppress this stemming, with the result that searches often return much more than you really want.

By default, **bibsearch** uses query-ranked searching: you type several words, and the search engine responds with a sorted list of bibliography entries that contain one or more of those words, in order of decreasing number of matches.

To switch to Boolean searching, which allows combination of words with Boolean AND (&), OR (|), and NOT (!) operators, issue the **mgquery**(1) command

```
.set query boolean
```

You may find that using

```
.set mode hilite
```

makes it easier to spot the matched strings in the output. However, the highlighting requires additional

control characters that contaminate any output directed to a file, so in such a case, you should turn it off, by, e.g.,

```
.set mode text
.set pager "cat >>/tmp/foo.log"
```

At the **bibsearch** development site, local users can see these files directly in the UNIX network file system, without having to launch **ftp**(1) or a Web browser: just change the prefix `ftp://ftp.math.utah.edu` to `/u/ftp` to create a local file name. This may be handy if you want to find related bibliography entries serendipitously, without knowing exactly what they contain.

For further information on searching with **mgquery**(1), consult its manual pages, or use its

```
.help
```

command.

Finally, to exit from **bibsearch**, use the **mgquery**(1)

```
.quit
```

command.

With each BIB_TE_X entry retrieved, **bibsearch** provides definitions of any BIB_TE_X strings used in the entry. This will likely result in many duplicate string definitions, but they are easily eliminated by a subsequent pass of the bibliography data through **bibsort**(1). Acknowledgement strings are not included in this output, because they are often large, and because few, if any, BIB_TE_X styles use them. You can find their definitions near the top of the original BIB_TE_X file identified at the start of each search result.

RELIABILITY OF BIBLIOGRAPHIC DATA

Although bibliographic data may be stored on a computer, it is ultimately generated by humans, and in the case of commercial databases, probably by low-paid, unskilled, offshore typists with no personal interest in accuracy. Worse, they are probably paid for speed and quantity, rather than quality.

Several years of experience with bibliographic data from many sources, both commercial and public, has amply demonstrated that if errors can be made, they will be! Among the common errors repeatedly seen are:

- Misspelled author names.
- Omitted accents in author names.
- Silent truncation of author lists, and in the case of one commercial database, the even worse practice of recording only the first two authors, plus the last!
- Replacement of authors after the first by the anonymous *et al.*. A colleague once quipped that if you hear of work credited to Jones et al., it means that Jones got the credit, but Al did the work!

Scientific document production is now mostly done electronically, so shortcuts of the past that were used primarily to reduce the tedium of manual typing of reference lists, such as omitting subsequent authors, abbreviating personal names or journal names, and eliding common digits of ending page numbers (983--7 instead of 983--987), should be abandoned. The document is now likely to be preserved in electronic form by the publisher, and those shortcuts interfere with subsequent reuse, and with searching.

A few journals already hyperlink citations to entries in reference lists, and inverting those hyperlinks produces answers to the very important question, "Who cited this article?", a question that up to now could be answered only through Science Citation Index, or its variants in other fields.

- Reduction of author personal names to initials, sometimes without final periods, or even separating spaces. In some languages, there are human names that require only a single letter, so a serious ambiguity is introduced by this reduction.
- Misordering of author names occurs far too often. Even publishers are sometimes inconsistent here between table of contents and article.
- Omission of spaces and hyphens, and loss of capitalization, in compound names. Science Citation Index reduces *Joop van der Vliet* to *J. Vandervliet*, *H. Fallah-Adl* to *H. Fallahadl*, and *S.*

McQueen to *S. Mcqueen*.

- Dropping of Jr.-like author name suffixes.
- Inconsistent handling of Chinese, Japanese, and Hungarian author names. In those languages, and possibly others, family names are given before the personal name. In the absence of bibliography markup of family names, it would be more consistent to place the family name last, e.g., *Paul Erdős*, instead of *Erdős Pal*, and *Bao-Wen Li* instead of *Li Bao-Wen*. A future version of BIB_T_E_X may have markup syntax to address this problem, while preserving the native name order, but until then, the common European name order should be adhered to.
- Incorrect abbreviation of Spanish names, which often have a paternal family name, followed by a maternal family name, often separated by *y* (and). Thus, *Javier Gomez Romero* should not be abbreviated as *J. G. Romero* or *Javier G. Romero*. For BIB_T_E_X use, enter such a name like this: *Javier {Gomez Romero}*.
- Confusion between editor and author. This is a frequent problem for journals that carry monthly columns with both a regular column editor, and an author varying by issue. Several IEEE journals do this. The BIB_T_E_X entry should list the author, not the editor.
- Errors in titles, sometimes making them unrecognizable. Part of this problem can be traced back to publishers, editors, and marketing people who prepare tables of contents with titles that only vaguely resemble the actual article titles. Some commercial databases seem to input their data almost exclusively from tables-of-contents listings, so the errors propagate.
- Changes of spelling between American and British English.

Databases should record the original data exactly, not convert it to the orthographic practices of the country where it happens to be stored!

When spelling errors occur in the original data, they can be marked with the time-honored Latin insertion [*sic*].

- Changing arabic numbers in titles to English, or vice-versa (a very common error in Science Citation Index).
- Errors from optical character recognition (OCR) when publication data is scanned in as bitmaps and converted to text by a computer program: *ri* becomes *n*, *ni* becomes *m*, *l* becomes *l*, and so on.
- Errors in year, volume, issue, month, and page numbers, and sometimes, even getting the journal completely wrong.
- Omission of publication month and, where relevant, day.
- Omission of final page numbers, even though it is required practice in many fields to include both initial and final page numbers in article citations.
- Off-by-one final page numbers, probably from faulty assumptions about whether new articles start on an odd-numbered page or not.
- Off-by-one initial page numbers from confusion over where an article starts.

Citations for the *Communications of the ACM* frequently exhibit this problem: since a design change in July 1990, articles often have background artwork on two facing pages, and huge and horridly scaled typefaces. The foreground and background colors and patterns often clash, to the point of being unreadable, and the weird layout of author names and titles leaves the reader thoroughly confused.

- Converting journal titles to all uppercase (a major failing of Science Citation Index).
- Dropping of punctuation: *O'Neill* becomes *ONeill* and *What's New?* becomes *Whats New*.
- Insertion of bogus hyphens between title words (a practice that is endemic in Science Citation Index).

The lack of support for accents, and for mathematical markup, is a huge problem for scientific bibliographic data. Only the American and European Mathematical Society databases, which store their data with \TeX markup, routinely include accents and proper math markup, although for historical reasons, the EMS databases typically use transliterations of German accented letters. Among all the commercial databases that I have used, the quality of entries in the AMS and EMS databases is notably better than all others, and they deserve a lot of credit for that.

The attempt at representing mathematics in many other databases cannot be termed other than bizarre and incomprehensible; they don't even attempt to do something sensible, like spell out the names of the symbols, but instead, substitute utter nonsense. Yet, mathematics is common in publication titles in the physical sciences.

Speakers of largely accent-free languages, like English, tend to be insensitive to the significance of accents in other languages. In many languages, accented letters are not just variants of the base letter, they are *completely different letters*, with different sounds and different alphabet positions. Just as English would become unreadable if the vowels were reduced from *aeiou* to, say, just *aiu*, other languages suffer similarly when accents are lost. Here is the previous sentence with that reduction:

Just as English would bicumi unriadabli if thi vuwils wiri riducid from *aiiuu* tu, say, just *aiu*, uthir languagis suffir similarly whin accints ari lust.

In view of these myriad deficiencies, what can be done?

The best way would be for commercial databases and publishers to do the job right in the first place!

Ultimately, commercial databases should get all of their data directly from publishers, instead of indirectly by retyping and/or OCR. Publishers have the original bibliographic information, and the capability to preserve that information, and make it available; at present, few do.

If publishers do not supply the correct information in rigorously-marked-up electronic form, then data should come from multiple sources and be merged in such a way that discrepancies are revealed. Commercial databases could have the data typed or scanned twice by independent contractors, for example, and then merge the results: discrepancies indicate errors. Researchers should enter bibliographic data directly from the original publications, or if those are unavailable, at least find multiple *independent* instances of citations to those publications.

The **bibjoin**(1) utility is a good example of software that can help a lot in this process. It uses a number of heuristics to maximize the data retained after a merge, and yet preserves conflicting information when it cannot decide which of multiple choices is the right one, leaving it up to a human editor to resolve discrepancies. Consult its manual pages for more information about how it operates.

However, merging of data from two sources is insufficient: both might have derived their data from inaccurate tables of contents.

For these reasons, the BibNet Project and TUG archives include data merged from several sources, and credit those sources in *bibsource* string values. In addition, for many recent journals, the data is derived directly by computer software from data at publisher World-Wide Web sites. For *some* of the larger scientific publishers, this has proved to be an acceptable approach, although confirmation from other sources is still desirable. Consequently, the **tug** collection in **bibsearch** will be found to be a substantially more reliable source than the other collections, which are mostly single source, and parts of them are very poorly and sloppily done.

The significant differences in quality is why separate databases are maintained for **bibsearch**, rather than merging all available data into one master database. While it is technologically feasible to do that, the user would then be unable to assess the quality of the results of a search, and the wheat could be buried in chaff.

DIRECTORIES AND FILES

`/usr/local/src/bibdata`

Installation-default **bibsearch** database directory search path.

`$HOME/.mgrc` User's personal configuration file for **mgquery**(1).

xxx/DESCRIPTION

Text file to provide brief descriptions of the contents of database **xxx**. The **bibsearch** --**help** option uses this file to prepare its database summary.

xxx/mgbuild

Temporary directory tree used by **mgbuild**(1) while creating a new database named **xxx**. After successful completion of the build, a fast directory renaming operation exchanges **xxx/mgbuild** and **xxx/mgdata** to make the newer one current. Since the old one remains available, just under a different directory name, open files in use by all running **bibsearch** or **mgquery**(1) jobs are unaffected.

Even when the old **xxx/mgbuild** tree is emptied immediately prior to the next scheduled database update, already-open files still remain available to running processes (but are no longer associated with a named file directory), so database updates never cause user search-process failure, but newly-added data will not be seen until the user search processes are restarted.

xxx/mgdata/bibfiles/STATS.*

Statistics files for database **xxx**. **bibsearch** uses the contents of these files in its welcome banner, with a report like this:

```
Last database update: Thu Feb  1 17:09:26 MST 2001
Bibliography entries:      278378
Bibliography lines:       5682256
Bibliography bytes:      216614068
```

xxx/mgdata/bibfiles/bibfiles.*

mg(1) binary index files for database **xxx**. These contain *all* of the data originally found in the BIB_TE_X files, but in a binary, compact, and heavily-indexed, form that can be accessed very quickly by **mgquery**(1). They typically take about 40% of the initial data size. The original BIB_TE_X files are never used by **mgquery**(1).

xxx/mgdata/mg_get

UNIX shell script helper program used by **mgbuild**(1) to create a marked-up data stream for indexing. It contains sections for handling data in various formats, and needs no modifications until data in a new format needs to be indexed.

xxx/mgdata/mg_get_bibtex.awk

Filter program used by **mg_get** to add **mgbuild**(1) markup to a BIB_TE_X data stream. It assumes that the data has been put into a standard format by **bibclean**(1), and generally needs no site-specific or user-specific modifications.

The database build time depends critically on the performance of this program. The current version is run with **mawk**(1), because that proved to be several times faster than with **awk**(1), **gawk**(1), **nawk**(1), and even with a highly compiler-optimized translation from **awk** to C with **awka**(1)! It now accounts for about half the total CPU time of the build of the **tug** database, for which **mgbuild**(1) processes about 7MB of clean BIB_TE_X data per minute, or 420MB per hour, on a 400MHz Sun UltraSPARC II file server.

ENVIRONMENT VARIABLES

BIBSEARCHPATH User-defined default **bibsearch** database directory search path. It overrides, or augments, the local system-wide **bibsearch** search path, and is in turn overridden by any command-line --**bibsearchpath** option. See that option's description in the **OPTIONS** section above for further details on how to define a directory search path.

DEBUG

If this is set to an arbitrary nonempty value, turn on debug tracing. This is almost the same as using the command-line --**debug** option, but makes the tracing take effect a little sooner.

SEE ALSO

agrep(1), awk(1), bibcheck(1), bibclean(1), bibdup(1), bibextract(1), bibjoin(1), biblabel(1), biblex(1), biborder(1), bibparse(1), bibsort(1), bibsplit(1), bibtex(1), bibunlex(1), citesub(1), emacs(1), glimpse(1), icon(1), mg(1), mgbuild(1), mgquery(1), perl(1), ruby(1).

AUTHOR

Nelson H. F. Beebe, Ph.D.
 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@computer.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

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

<ftp://ftp.math.utah.edu/pub/mg/mg-1.3x>
<http://www.math.utah.edu/pub/mg/mg-1.3x>

in the files

bibsearch-x.yy.jar
bibsearch-x.yy.tar.gz
bibsearch-x.yy.zip
bibsearch-x.yy.zoo

where *x.yy* is the current version. Each of the popular archive format unpacks into an identical distribution tree.

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 *bibsearch* 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

```
#####
#####
#####
###                                     ###
```

```

###          bibsearch: search BibTeX bibliography files          ###
###
###          Copyright (C) 1997, 2000 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                                             ###
#####
#####
#####

```