BibUpdate

Andrew Mathas

Version 2.1.0

Automatically update bibtex files

Usage:

```
bibupdate [-h|-H] [-a] [-x] [-c] [-k] [-i FIELDS] [-l LOG]
          [-m | -M | -n] [-o] [-q] [--threads N] [-w LEN]
          [--deduplicate] [--remove-duplicate-papers] [--rename-duplicate-keys]
          [--list-duplicates] [--list-papers] [--list-keys]
          [-t FILE.tex [FILE.tex ...]]
          [--output FILE.bib] bibtexfile [outputfile]
bibupdate --merge [options] FILE.bib FILE.bib [FILE.bib ...]
```

This is a command line tool for updating the entries in a BibTeX file using the tools provided by the American Mathematical Society: mrlookup, MathSciNet and MRef. By default bibupdate uses mrlookup to update `article`, `book`, `inproceedings` and `incollection` entries when they do not have an `mrnumber` field. Entries are not updated if they already have an `mrnumber` field, even an empty one, unless `--all` or `--check` is used. The `-x` option makes bibupdate look for each paper in the arXiv and add `eprint` and `archiveprefix` fields for all the papers that it finds. In addition, bibupdate is able to merge bibliography files, removing duplicate papers and keys in the process and, optionally, updating citations to renamed duplicate keys in LaTeX files.

By default, bibupdate does not change your original file. Instead, it writes the updated entries to a new file, with the filename given by adding an `updated_` prefix, so *file.bib* becomes *updated_file.bib*. To update the bib file in place, use `--overwrite`, or give a output filename. Existing output files are backed up with `.bak` extensions, but `--overwrite` should still be used with care.

This should be used with care because you can potentially lose data.

Options by task:

Updating and checking

```
-m, --mrlookup        use mrlookup to update bibtex entries (default)
-M, --mathscinet      use mathscinet to update bibtex entries (less flexible)
-n, --mref            use mref to update bibtex entries
-a, --all             update or validate ALL BibTeX entries
-x, --arxiv           add arXiv eprint fields to entries
-c, --check           check all entries and report duplicates without writing a file
```

Duplicate listing

```
--list-duplicates     list duplicate papers and keys without writing a file
--list-papers         list duplicate papers only
--list-keys           list repeated citation keys only
```

Duplicate cleanup

```
--deduplicate         remove duplicate papers and rename repeated keys
--remove-duplicate-papers
                      remove duplicate papers, keeping the first
--rename-duplicate-keys
                      rename repeated keys using authors and year
--merge               merge two or more .bib files and clean up duplicates
-t, --texfile FILE.tex [FILE.tex ...]
                      update citations in listed TeX files during cleanup
```

Output and formatting

```
--output FILE.bib     write the bibliography to this file
-o, --overwrite       update/overwrite the bibtex file (use with care)
-i FIELDS, --ignored-fields FIELDS
```

```
                      space-separated list of bibtex fields to ignore
-k, --keep-fonts      do NOT replace fonts \Bbb, \germ and \scr in titles
-w LEN, --wrap LEN    wrap bibtex fields to specified width
--threads N           number of concurrent AMS lookups (default: 1)
```

Reports and logging

```
-q, --quiet           brief updates and missed lookups
-qq                   only print errors
-l LOG, --log LOG     log messages to specified file (defaults to stdout)
```

Help and version

```
-h, --help            show grouped options and exit
-H, --Help            print full program description
```

Note: As described below, you should check the updated `.bib` file for errors before deleting the original BibTeX file.

BibTeX is widely used by the LaTeX community to maintain publication databases. This script attempts to add missing fields to the papers in a BibTeX database file by querying mrlookup and getting the missing publication data from there. This is not completely routine because an mrlookup search requires either the names of the authors, or the title of the article, and both of these can have non-standard representations. If the article is already published, then it is also possible to use the publication year and its page numbers as part of the search. To search on mrlookup we:

- use the first author's surname (accents and compound surnames can be problematic)
- use numeric page ranges, or a single numeric page, when available
- include the year for books, or when a numeric first page is available
- use the cleaned title when a numeric first page is not available

The author query uses only the first author's surname because mrlookup does not accept multiple author names joined by 'and'. Braces used to protect capitals in titles (for example `{Kac-Moody}`) are removed when searching and matching, with the title stored in the BibTeX file retaining its braces.

If there is a unique match from mrlookup, then bibupdate updates the entry in the BibTeX file using the field values returned by an mrlookup search, except for the citation key and any fields excluded by `--ignored-fields`. Any fields that are not returned by mrlookup, such as `eprint`, are retained. By default, a message is printed whenever an entry in the database is changed. If the title of the retrieved paper does not (fuzzily) match that of the original article, then the entry is NOT updated and a warning message is printed.

Although some care is taken to ensure that the new BibTeX entries correspond to the original entry in the BibTeX file, there is always a (small?) chance the new entry corresponds to a completely different paper. In my experience this happens rarely, and mostly with unpublished manuscripts. In any case, before you delete your original BibTeX file *you are strongly advised to check the updated BibTeX file carefully for errors!*

By default, the program prints a compact listing of any updated paper in the BibTeX file, including any added arXiv fields. This can generate a lot of output the first time bibupdate is run on large files, later runs produce much less output.

As bibupdate calls mrlookup, this program will only be useful if you have papers in your database that are listed in MathSciNet. As described below it is also possible to call MathSciNet directly, however, this is less flexible because the `mrnumber` field for each paper is required.

I wrote this script because I wanted to automatically update the papers in my BibTeX files when they were published, so that the file has up-to-date information on all of the papers in the file, including journal names, arXiv links, and DOIs for each paper. As a bonus the script helped me to correct many minor errors that had crept into my BibTeX file over the years (for example, incorrect page numbers and publication years).

Options and defaults

By default, bibupdate uses mrlookup, sending one AMS lookup request at a time. It prints a compact listing for each entry in the file that is updated, writing the updated BibTeX file to `updated_file.bib` if the original input file was `file.bib`.

Updating and checking

`-m, --mrlookup`

Use mrlookup to update BibTeX entries (default)

`-M, --mathscinet`

Use MathSciNet to update BibTeX entries

`-n, --mref`

Use MRef to update BibTeX entries

By default mrlookup is used. Its main advantage is that it has a more flexible search mechanism using author, title, year and page numbers, so it can find papers even when they do not yet have an `mrnumber`.

With `-M`, the entry is looked up using its `mrnumber`. Entries with missing, empty or non-numeric MR numbers are skipped, including with `--all` and `--check`.

With `-n`, the full BibTeX entry is handed to MRef as a free-form reference. The best matching record is then checked against the original title. Like mrlookup, MRef normally checks only `article`, `book`, `inproceedings` and `incollection` entries without an `mrnumber` field. Use `--all` to check all recognised entry types.

`-a, --all`

Update or validate ALL BibTeX entries

`-x, --arxiv`

Update the arXiv entry for each paper

When the paper is found, add `eprint` and `archiveprefix` fields to the BibTeX entry:

```
eprint = {2211.07069},
archiveprefix = {arXiv}
```

bibupdate searches the arXiv using the significant words in the title. A paper returned by the arXiv is counted as a match only when there is a unique paper with a sufficiently good (fuzzy) match. The `-x` option can be combined with any of the `-m`, `-M` or `-n` lookup tools. Entries with an existing `eprint` field, even an empty one, are skipped unless `-a` or `--check` is used. Entries without a usable title cannot be searched.

As requested by the arXiv, bibupdate waits at least 3 seconds after each completed arXiv request before sending the next one. Updating a large database with `-x` can take a long time.

`-c, --check`

Check all entries and report duplicates, printing compact listings for each database entry that would be changed. No output file is written. This option implies `--all`.

`--threads N`

Number of concurrent AMS lookups

Use `--threads N` to run `N` searches in parallel. arXiv lookups remain sequential and keep their delay between requests, regardless of the thread count.

Duplicate listing

`--list-duplicates`

List any duplicated papers, or citation keys, in the BibTeX database.

`--list-papers`

List duplicate papers only, including possible title-and-author matches

Two entries are treated as the same paper if they share a DOI, an MR number or an arXiv identifier. As a fallback, entries whose normalized titles and author lists (surnames and first initials) agree are reported as *possible* duplicates. Entries linked through a chain of such matches are grouped together. The same rules are used by `--check`, `--remove-duplicate-papers` and `--deduplicate`.

`--list-keys`

List repeated citation keys only

Duplicate cleanup

`--deduplicate`

Remove duplicate papers and rename repeated citation keys

This offline cleanup first removes repeated papers, then renames repeated keys among the remaining entries. As usual, the updated file is written to `updated_file.bib` unless `--overwrite` is used. Existing fields and formatting are preserved; missing fields are added from later copies. Ignored-field, font-conversion and wrapping options do not apply to cleanup. Cleanup cannot be combined with listing options or `--check`, and the lookup options (`-m`, `-M`, `-n`, `-a` and `-x`) are ignored.

`--remove-duplicate-papers`

Remove repeated papers, keeping the first occurrence in file order

Papers are matched as described under `--list-papers`. Title-and-author matching is heuristic, so inspect the output of `--list-papers` before removing entries.

The first entry is the base. Missing fields are copied from later entries in file order. Existing fields, including empty fields, are kept; if several copies supply the same missing field, the first value wins. Field expressions and macros are copied without expansion. Each removal reports the removed key and the retained key, with source lines. Literal `crossref`, `xref`, `related`, `entryset` and `xdata` references inside the bibliography are updated where the mapping is unambiguous. Use `-t` to update citations in external LaTeX files; otherwise these must be updated manually.

`--rename-duplicate-keys`

Rename repeated keys without deleting papers. The first occurrence of each key is kept. Duplicated key names are replaced with a key formed from the authors' surnames and the year, for example `AcharRiche2019`. Editors' names are used if there are no authors. If necessary, a suffix of `a`, `b`, ..., `z`, `aa`, etc. is added to ensure that the key names are unique.

Using `--remove-duplicate-papers` together with `--rename-duplicate-keys` is equivalent to `--deduplicate`.

`--merge`

Merge two or more BibTeX files, removing duplicate papers and keys

The input files must have a `.bib` extension and are processed in the order given. When removing duplicates, the rules used by `--deduplicate` apply: the first copy of each paper and key is kept, and any repeated keys are renamed using the authors' names and the publication year. Fields missing from the first copy are added from later entries, with earlier values having precedence. Retained entries, comments, preambles and string macros keep their original formatting; macros are not expanded for duplicate matching.

Merging is offline and cannot be combined with `--check` or duplicate listing options.

By default, the merged BibTeX file is written to `merged_first.bib` if the first BibTeX file is called `first.bib`. Use `--output` to write to a different filename, or `--overwrite` to overwrite the first BibTeX file, which will be backed up with a `.bak` extension.

```
bibupdate --merge first.bib second.bib --output combined.bib
bibupdate --merge first.bib second.bib -t paper.tex --output combined.bib
```

Use `-t`/`--texfile` to update unambiguous citations in associated LaTeX files, which are backed up with `.bak` extensions.

`-t, --texfile FILE.tex [FILE.tex ...]`

Update citation keys in the listed LaTeX files when cleaning duplicate entries

This option requires `--merge`, `--deduplicate`, `--remove-duplicate-papers` or `--rename-duplicate-keys`. The `.tex` and `.bib` extensions distinguish the TeX files from the bibliography and optional output file, for example:

```
bibupdate --deduplicate -t chapter.tex appendix.tex papers.bib
```

Only unambiguous citation key mappings are applied to the LaTeX files. If different papers share a key, citations of that key in the LaTeX files are left unchanged and a warning is printed. The later papers are given new unique keys in the merged BibTeX file.

Standard literal citation commands such as `\cite`, `\citep`, `\nocite`, `\parencite` and `\textcites` are supported, including optional notes and comma-separated key lists. Comments, verbatim text and other commands are preserved. Custom citation macros are not expanded, and included files are only processed if explicitly listed.

Before changing a TeX file, its original contents are saved with a `.bak` extension, for example `chapter.tex.bak`. Unchanged files are not rewritten or backed up. Existing backups of changed files are replaced.

Output and formatting

`--output FILE.bib`

Write the bibliography to the specified file

Use this option to specify the name of the new BibTeX output file. An existing non-empty output file is first backed up with a `.bak` extension. This option cannot be combined with `--overwrite`, or with an output file given as a second positional argument.

`-o, --overwrite`

Overwrite the existing BibTeX file with the updated version. The original file is first backed up with a `.bak` extension. This option cannot be combined with `--output`.

`-i FIELDS, --ignored-fields=FIELDS`

A space-separated list of BibTeX fields to omit when writing the updated BibTeX file

By default bibupdate removes the following fields from each BibTeX entry:

- `coden`
- `mrreviewer`

The `-i` option replaces this default list. Repeated `-i` options are combined; field names are case-insensitive and separated by spaces. An empty string keeps every field:

```
bibupdate -i "coden fjournal" file.bib   # ignore coden and fjournal
bibupdate -i coden -i fjournal file.bib  # ignore coden and fjournal
bibupdate -i "" file.bib                 # do not ignore any fields
```

`-k, --keep-fonts`

Use `-k` to retain legacy LaTeX font names in titles. By default, the following font substitutions are automatically applied:

- `\Bbb X` --> `\mathbb{X}`
- `\scr X` --> `\mathcal{X}`
- `\germ X` --> `\mathfrak{X}`

This applies both to input titles and to titles retrieved from AMS.

`-w LEN, --wrap LEN`

Wrap BibTeX fields to specified width

By default, lines are not wrapped (a wrap length of `0`). A positive wrap length `LEN` wraps each field value to that width, with a tab before continuation lines.

Reports and logging

`-q, --quiet`

Print fewer messages

By default, bibupdate prints a compact reference for each paper that is updated, and a warning for each paper for which the AMS lookup fails. With `-q` (or `--quiet`) only a one-line message is printed for updated papers, together with the warnings for failed lookups. With `-qq` only errors are printed. Failed lookups are printed in red when the output is a terminal.

`-l LOG, --log LOG`

Log output to file (defaults to stdout)

By default, reports and warnings go to standard output. Use `-l` to write them to a file. Existing log files are overwritten. Fatal errors are always printed to standard error.

Help and version

`-h, --help`

Print a short list of command-line options

`-H, --Help`

Print the full-help message

Known issues

- **BibTeX syntax:** the parser uses regular expressions rather than a full BibTeX grammar. Use brace-delimited entries with one field per line and braced values or simple words. Multiple fields on one line, quoted values, `#` concatenation, string-macro expansion, and fields containing `@` or `={...}` can be misread or omitted. Inspect the output carefully when using these forms. Duplicate listing and cleanup use a source scanner that preserves quoted and nested values, comments and macros; macros are not expanded for duplicate matching.
- **Title matching:** capitalization braces and apostrophes are removed before searching and comparing. Mathematics is also removed, which can make distinct titles look similar. TeX commands and accents are not fully decoded, so equivalent spellings can still fail to match. A unique fuzzy title match is a heuristic, not a guarantee that a paper has been correctly identified.
- **Author names:** mrlookup uses only the first author's surname. A compound surname is preserved in `surname, given names` form; without a comma, only the last word is used. Accents, name particles and alternative spellings can therefore affect searches.
- **Page numbers:** numeric ranges and single numeric pages are supported. Article identifiers and prefixed page numbers are not handled reliably. When only non-numeric page numbers are given, the title is used instead. For non-book entries the year is then omitted from the query.
- **Ambiguous or missing results:** an AMS entry is updated only when there is a unique match. BibTeX entries are not changed when no matches, or multiple matches, are found. Entries without a usable title are more likely to receive an incorrect match. By default, skipped entries, such as those with an existing `mrnumber`, are never changed.
- **arXiv searches:** searches use significant title words and examine at most ten results. Titles dominated by mathematics or short words are more likely to receive false matches. As with mrlookup searches, the arXiv fields for a paper are updated only if there is a unique match.

- **Network services:** connections can fail, time out or return unexpected data. Requests use a 30-second timeout. Connection errors and timeouts produce warnings, and processing continues. A missed lookup does not prove that the paper is absent from the database.

Feature requests and bug reports

Please make any feature requests and bug reports on the BibUpdate github page:

https://github.com/AndrewMathas/BibUpdate/issues.

Please provide as much detail as possible when submitting requests.

Bug reports must be accompanied by a *minimal working example*, which is a minimal BibTeX entry that demonstrates the problem.

Licence

Andrew Mathas

GNU General Public License, Version 3, 29 June 2007

bibupdate Version 2.1.0. Copyright (C) 2012-2026

This program is free software: you can redistribute it and/or modify it under the terms of the GNU General Public License (GPL) as published by the Free Software Foundation, either version 3 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.